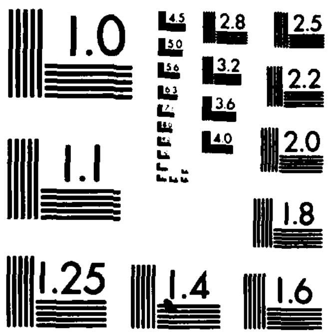
AD-A172 362 OPTICAL PROCESSING(U) DAYTON UNIV OH RESEARCH INST 1/2
A C WALKER ET AL. JUN 86 N00014-85-K-0479

UNCLASSIFIED F/G 20/6 NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AD-A172 362

Contract No. N00014-85-K-0479

Title: Optical Processing

FINAL REPORT

June 1986

Heriot-Watt University

University of Dayton Research Institute Subcontract Order No. RI-43109

ONR Main Contract No. N00014-85-K-0479

Authors: A.C. Walker and W. Taylor

86 7 18 144

'DEVELOPMENT OF OPTICAL BISTABLE SEMICONDUCTOR LOGIC ELEMENTS AND ARRAYS FOR HIGH-SPEED COMPUTING APPLICATIONS'

Physics Department, Heriot-Watt University

Riccarton, Edinburgh, Scotland, UK.

1. INTRODUCTION

This report summarises progress over the five month period of Jan-May 1986. Much of the work completed has been of a preparatory nature - laying the groundwork for the future programme. It has included the purchase and commissioning of experimental hardware and the execution of a number of preliminary experiments. The short time-scale should be noted in the context of the originally proposed Statement of Work, which envisaged a period of a full year for this study. The following reviews progress under those headings listed in the initial proposal.

2. MAIN AREAS OF WORK

2.1 Multiple Focus Optical Elements

The major part of this work is directed towards development of large arrays of optically bistable gates for parallel processing applications. An important requirement of such arrays is an efficient system of illumination. Assuming a clear separation between each gate and its neighbours, then it is necessary to break the laser beam, which acts as the power input to the array, into a corresponding pattern of focussed beamlets. Holographic techniques are particularly suited to this purpose. In addition, we anticipate that holographic interconnects will play an important role in combining numbers of optical gate arrays to form parallel optical digital processors. With these requirements in mind we have assembled a facility

dedicated to the production of holographic optical elements (HOEs) for specific use in optical digital circuits.

To ensure high efficiency exploitation of available laser power we have chosen to use Dichromated Gelatin (DCG) as the (volume) emulsion, as this readily yields nearly 100% diffraction efficiencies. The need for reproducible, predictable results when fabricating HOEs using DCG dictates the use of an environmentally controlled clean-room. Such a facility has now been commissioned and a vibration-isolated optical table plus line-narrowed argon-ion laser (Innova 90-6) also installed. The room has a temperature specification of 21 ± 2°C and a relative humidity specification of 35 ± 5%. Also included in the room is a wet-processing area inside an air-extract station. This permits the full preparation, exposure and processing cycle to be completed under controlled conditions of temperature and humidity. The laser can also be used for characterisation of the final HOEs, including the illumination of optical gate arrays and their assessment.

2.1.1 HOE Array Production

Figure 1a shows the layout that we typically use to make the holographic lenslets. The point-source 'object' beam is normally incident on the sensitised plate, while the reference beam lies at an angle of 40° from normal. This geometry ensures that, on reconstruction (see Fig. 1b), the transmitted -1 diffraction order lies near the plane of the hologram itself and therefore tends to be suppressed. Together with the volume nature of the hologram this permits high efficiencies to be obtained for light diffracted into the desired +1 order.

Each lenslet is produced by a typically ~ 0.2 sec exposure and the plate is then stepped sideways ready for exposure of the next element in the array.

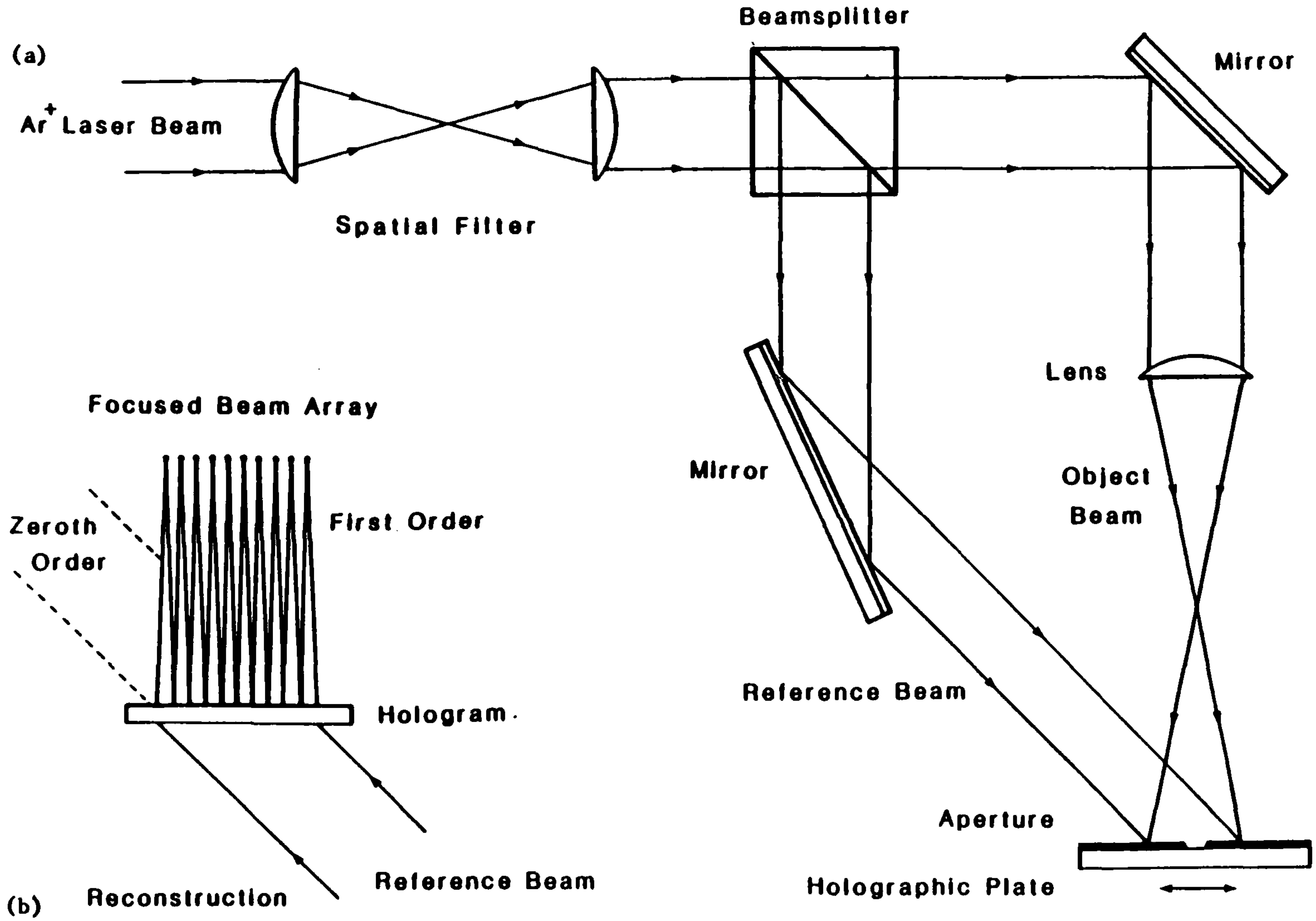


Figure 1 Experimental Arrangement for Construction Of Holographic 2-D Array Generator

We are currently setting up an automated step and repeat exposure system to permit convenient production of large lenslet arrays.

A range of techniques for preparing the DCG plate has been investigated. At present, we are obtaining good results using commercially prepared, 15 μm thick, gelatin layers which we then sensitise with ammonium dichromate.

2.1.2 Results

Single HOE lenslets have been produced with efficiencies > 95% and large arrays of up to 100 elements constructed in preliminary tests. The essential requirements for our application are high efficiency and uniformity across the array. Thus more recently we have been concentrating on the production of carefully controlled HOE arrays with the aim of optimising reproducibility.

An example of our current capability is a square 5 x 5 HOE array consisting of 2 mm diameter 60 mm focal-length DCG lenslets on 2 mm-spaced centres. Each element has the same 80% diffraction efficiency, within a measurement accuracy of 2%. They produce focal spots of ~ 70 μm diameter - within a factor 2 of the diffraction limit. The holographic medium is protected by an optically-cemented cover-slip and neither this nor the glass substrate are anti-reflection coated. Taking into account the illumination angle, we can conclude that a suitable a-r coating would raise the efficiency with which light is diffracted into the desired order to 92%.

2.1.3 Future Work on HOEs

Larger arrays will be easily fabricated once the automated step-and-repeat exposure system is fully commissioned. The ultimate limit upon the number of elements in the array will depend on the maximum acceptable overall dimension and the required focal-spot sizes. Smaller lenslets may be used provided shorter focal-lengths are employed to preserve

the necessary numerical aperture. This would lead to either a close-coupled geometry, in which the HOE array and optical gate array are spaced a short distance apart, or a decoupled geometry in which the bias beamlet array, generated by the HOEs, is imaged by an intermediate, longer focal-length, high N.A. lens onto the gate array from some distance [Recent Publications: 19].

An alternative approach is to produce either partially or fully overlapped HOEs. These could exploit most (or all) of the exposed aperture in generating each focal spot in the array and thus avoid a trade-off between array numbers and focal spot size diffraction limits. They could also have the advantage of not requiring a spatially uniform illumination. We have initiated a study of two possible techniques for making such 'fan-out' HOEs. The first technique is to make repeated exposures on the same area of the DCG plate while stepping the point-source 'object' across the required array grid. Preliminary experiments have shown up the expected problem of balancing up each exposure to ensure simultaneously equal and high diffraction efficiencies for each point in the array. We note that low numbers of overlapped exposures in DCG can yield high efficiencies [1,2]. The second technique treats the array as a single entity, to be recorded in the same way that a hologram of any object is made. In this case the array can be initially generated inefficiently, e.g. using masks, beamsplitters, computer generated holograms, etc., and then recorded for high efficiency reconstruction. We have performed some preliminary experiments of this type using the 25 element array described earlier to generate an object pattern. Further work needs to be done to eliminate the interference between components of the array that causes ghost images and brightness variations in the reconstructed image.

2.2 Packing Density Limits

A possible advantage of optically bistable gate arrays is that physical definition of each gate may not be required. Instead a uniform optically nonlinear device could be illuminated by a multiple focus array generator (as described in the previous section) and the gate distribution determined solely by the illumination pattern. The limiting factor in this approach is the degree of diffusive cross-talk between gates, which will determine the maximum packing densities.

With the simple thermally-based optically nonlinear devices, that we are currently exploiting to develop optical digital processing concepts, transverse thermal diffusion is the mechanism responsible for gate cross-talk. Some initial experiments have been carried out to quantify this effect.

2.2.1 Results

A simple two beam experiment has been performed using a uniform nonlinear thin-film multilayer deposited on a 2 mm-thick glass substrate. A copper heat-sink was clamped around the edges of a ~ 1 cm^2 exposed area. The two beams were focussed upon well separated positions on the device and adjusted so that each optical gate was held in the bistable region. The power in one beam was then raised slightly so as to induce switching and the second gate monitored. It was found that if the second gate was initially within 5% of its own switch power, switching was induced by the action of the first gate if they were within a distance of ~ 2 mm of each other.

By heat sinking across the back of the device, e.g. using a sapphire plate, longitudinal heat-flow could be enhanced and consequently the transverse cross-talk reduced. In this way we have lowered the requirement

for 2 mm separation, just described, down to ~ 0.75 mm. Further improvements are clearly necessary.

2.2.2 Future Work

It would appear that transverse cross-talk is a serious problem in thermally based bistable devices with uniform construction. Similar cross-talk effects have also been seen in optically bistable devices based on electronic-transition nonlinearities, due to carrier diffusion [Recent Publications: 9,10]. We conclude that physical pixellation of these devices is required if arrays containing 10^4 to 10^6 gates cm^{-2} are to be realised.

Some preliminary experiments on relatively crude devices incorporating both mechanically-machined and laser-cut pixellation have shown that 100 μm pixels on 200 μm centre spacing can act as independent gates. Exploitation of more sophisticated etching techniques can be anticipated to provide 10-100 μm gate separation, the required gate densities and simultaneously, as a result of inhibiting diffusion, lower operating powers. A programme of further work in this area is being proposed based on designs such as that in Figure 2.

2.3 Thermal Power Dissipation

A crucial aspect of any nonlinear optical gate array, as with electronic gate arrays, is the removal of the thermal power dissipated in the device. This is equally the case for devices based on thermal nonlinearities and those based on electronic-transition nonlinearities. Both involve real excitation of the medium and subsequent relaxation to thermal energy (dominantly).

2.3.1 Results

Less experimental work has been done in this area as we still await the production of arrays sufficiently large to yield serious cooling problems.

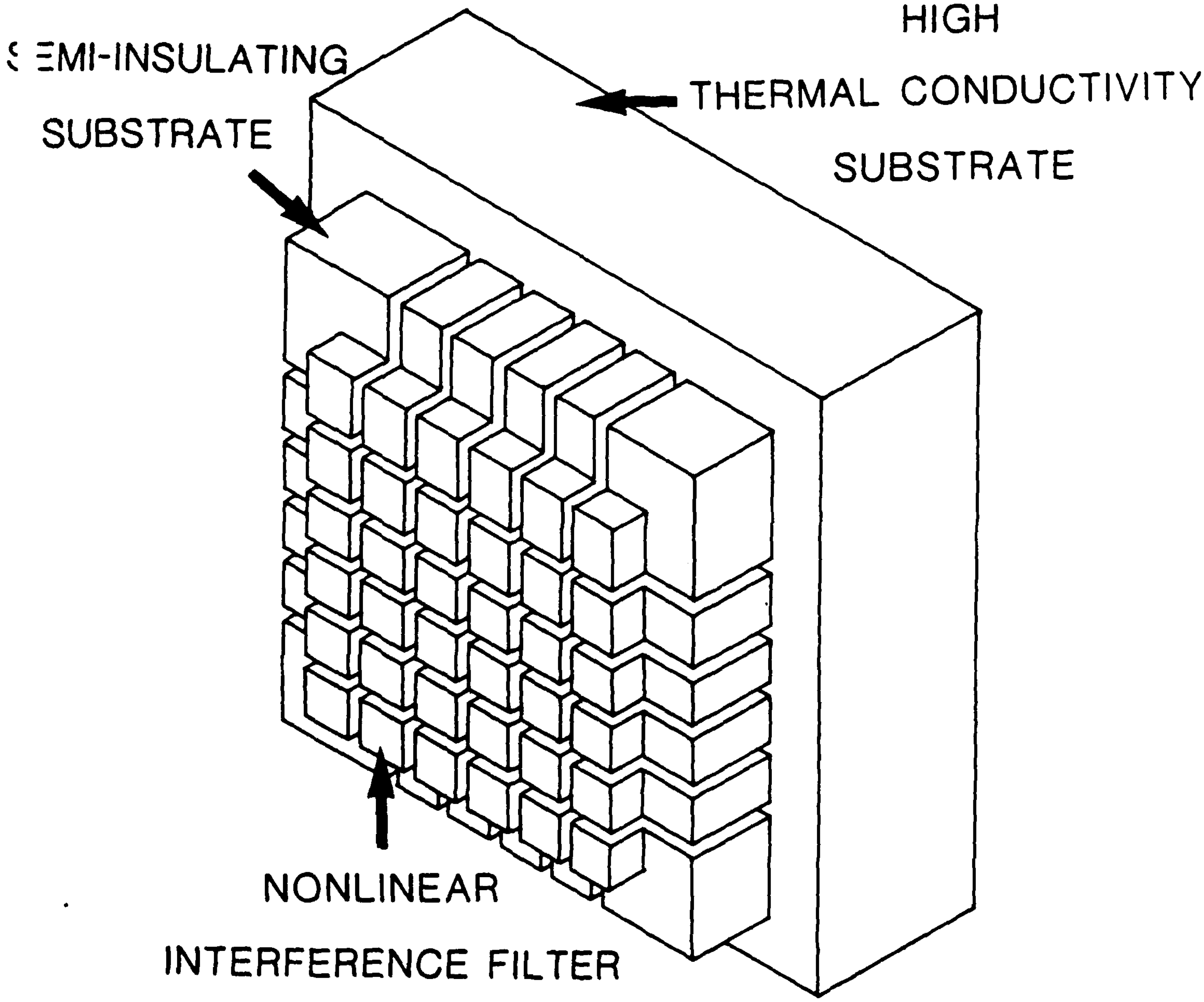


Figure 2 Design for a low-crosstalk optical logic array

In the meantime we have performed some experiments using crystalline substrates as heat-sinks for our multilayer thin-film bistable devices. In addition to their high thermal conductivity these materials can also be optically transparent and therefore permit close coupling to a 2D array without inhibiting transmission. A structure currently being studied (see Figure 2) is one in which the multilayer is deposited on a thin glass-substrate which, in turn, is cemented to a sapphire disc. The glass acts as an insulating buffer - permitting the necessary temperature changes to be induced in the thin-film. Variation of its thickness can be used as a means of adjusting the balance between device sensitivity and speed.

2.3.2 Future Work

Already under development is a convenient heat control packaging for these gate arrays. This will include active temperature control of the array and will act as the final heat dissipation component. The temperature sensitivity of the nonlinear etalons being employed to make optically bistable gates can be turned to advantage by also using this temperature control facility to adjust the initial detuning from cavity resonance to the precisely required value.

2.4 Prototype Gate Arrays

The 25 element HOE array, described in section 2.1.2 has been used to illuminate one of our nonlinear multilayer thin-film bistable devices. This allowed us to successfully demonstrate a 0.7 cm^2, 5 x 5 array of optically bistable switches.

2.4.1 Results

Using 3 watts of 514 nm radiation all of the 25 gates could be simultaneously switched into their high transmission states. Individual-

gates could be switched by a transient external signal input from their low to high transmission states while held in a bistable region. Switching times were of the order ~ 1 ms. No attempt was made to minimise either operating power or switch time.

As already stated the HOE multiple focus array generator had good uniformity (< 2% efficiency variation between elements). The nonlinear thin-film device is also very uniform and was designed for normal incidence bistability - avoiding variations due to focal depth limitations. Thus identical gate responses could be expected from this array. In practice this has not yet been tested because of the requirement for uniform input illumination of the HOE array. Expanding the central region of the available Gaussian profile laser beam to provide such a constant irradiance over the array area is too inefficient an approach. We are currently fabricating a holographic component to provide the required flat profile and also assessing a variety of other possible techniques.

It is interesting to note that no special heat-sinking was employed in this prototype demonstration. The nonlinear multilayer was deposited directly onto a 2 mm-thick glass substrate which was loosely supported around its edge by a copper holder. No gross long-term heating effects were observed despite over 1 watt power dissipation.

3. GENERAL TOPICS

The following areas were also listed in our original Statement of Work as relevant to the development of optically bistable gate arrays and to be therefore kept under review.

3.1 Switch Energy Reduction

Clearly for large gate arrays to be useful each element must work with a minimum input power and minimum switch time. In general for a particular type of device there is a trade-off between these two parameters such that their product, i.e. the switching energy, E_s, remains roughly constant. This can be seen in our thin-film thermal-mechanism devices when the substrate conductivity, K_s, is varied. Because the switch power : $P_s \propto K_s$, and the switch time : $\tau_s \propto K_s^{-1}$, it follows that although each can be adjusted by changing K_s their product remains constant.

However, as with most logic switches, switching energies can be reduced by making the gates physically smaller. This is particularly the case with our unpixellated thin-film devices. As the illuminating focal spot is reduced in size the switch power has been found to fall in proportion to its diameter. More dramatically, the switch time scales roughly as the square of the diameter. Thus the energy is reduced in proportion to the cube of the linear dimension of each gate. Typically for a 10 μm gate size our current devices have ~ 100 nJ switch energy. However, it is clear that this figure can be considerably reduced by a number of optimisation techniques[Recent Publications: 16, 18] currently under study: possibly down to ~ 1 nJ for a 10 μm x 10 μm isolated pixel. Still smaller gate sizes, assuming a physically pixellated device, could give E_s scaling of ~ 10 pJ/μm^2. Future

experiments are planned to assess the practical potential of these new design concepts. (A crucial requirement in this optimisation is the need for physical pixellation, as discussed in section 2.2.2).

These projected switch energies imply gate arrays could be constructed from nonlinear devices based on thermal effects with the capability of 10^9 - 10^{10} switch operations sec^{-1} $watt^{-1}$. Although this does not match some of the projected responses of other faster devices it is sufficiently large to permit serious experimental assessment of a variety of all-optical parallel processing concepts. Furthermore, and crucially, these relatively simple devices are far more practical than many current alternatives.

(1) They operate at room temperature,

(2) They work at a visible wavelength corresponding to a high power cw laser,

(3) They can be easily fabricated as large area uniform devices.

It is for these reasons that it is most important to pursue the development of this type of device and exploit their suitability in the continuing development of all-optical digital computing circuits. At the same time we are closely following and indeed contributing to the development of alternative devices that in the long run may prove to be suitable substitutes for them [Recent Publications: 3,4,5,9,11,17]. Significantly, many of the technological and architectural aspects are independent of the actual gate mechanism and will be relevant to any type of optically bistable logic array.

3.2 Materials and Fabrication Optimisation

As indicated in the conclusion of the previous section (3.1) we have under continuing review alternative materials for fabricating optically

bistable devices - both of the same, thermal-mechanism, variety as our thin-film multilayer devices and of the other, e.g. electronic transition mechanism, type of device such as our InSb switches. In addition fabrication techniques are also being assessed within a number of areas. For example, the problem of making high quality solid Fabry-Perot etalons of thickness $\sim 10\ \mu m$ is being studied. (Typically, we use currently either $\sim 100\ \mu m$ thick bulk material or $\sim 1\ \mu m$ films). Two possible approaches to preparing intermediate thickness devices are being investigated: (i) polishing of bulk crystalline material down to the required thickness (limited by material weakness and flexibility), and (ii) direct growth of thick layers of the dimensions needed (in some cases limited by stress build-up within thick films). This problem is relevant to all bistable optical gates relying on nonlinear dispersive phenomena in interference structures and a major part of our future development work is directed in this area.

Another point of current concern is the long-term structural stability of the materials being used, particularly when actually being operated as optical gates. We have found with our thin-film multilayer devices that long-term hold in a state of high transmission and intense illumination causes structural changes sufficient to seriously alter the input/output characteristic of the devices. Experiments indicate that this effect is strongly dependent upon the internal irradiance levels and suggest that photo-structural changes are being induced. Such phenomena are known to occur within the micro-crystalline layers generated by thermal evaporation techniques. For these devices to continue as the easy-to-use work horses of optical parallel digital processing research it is essential for this structural stability to be improved. To date we have shown that the detailed design of the multilayer can be optimised to reduce this effect. In the

future, however, we plan to exploit alternative growth techniques to ensure material with greater intrinsic structural stability is used at the outset.

3.3 Development of New Architectures

We have continued to develop architectural concepts of relevance to the immediate exploitation of our available optical gate arrays within proof-of-principle optical digital processing demonstrations [Recent Publications: 2,19]. In particular the 'lock and clock' approach is being pursued as a means of constructing an elementary classical finite-state-machine. This system, currently nearing completion, is a development of the all-optical digital circuits successfully operated previously (Recent Publications: 1,20]. It has initially been constructed around simple 3-element arrays - to demonstrate a minimum level of parallel'sm. The much larger gate arrays which we have started to develop under this contract will permit much higher degrees of parallelism and provide an opportunity to perform genuine optical digital processing of parallel input data.

4. TECHNICAL BUDGET SUMMARY

4.1 Equipment

The major items of equipment purchased included a Coherent Model 90-6 Argon Ion laser, a Photon Control optical table, optical bench components and accessories, a microcomputer system and environmental control equipment. This has been used to commission and equip a new clean-room facility with filtered air and temperature and humidity control. Additional works for this facility (plumbing, electrical and other services) were funded from overheads. The laboratory is now commissioned and has already been used to fabricate and operate a 5 x 5 zinc selenide holographically illuminated optical logic array as described in section 2. The controlled conditions of this laboratory are essential for the successful reproducible processing of holographic optical elements. A computer controlled system is used for accurate exposure of the holographic arrays.

4.2 Travel Funds

Owing to a six-month delay in awarding the contract and a reduction in the expected period of funding, visits to certain key conferences and meetings in the U.S. were missed and consequently the allotted travel funds were not fully utilised. Permission was sought to use any remaining funds for the purchase of additional necessary equipment; the main item acquired in this way was a helium-cadmium laser which has already been used to successfully demonstrate differential wavelength addressing of a logic array. This laser provides new source wavelengths in addition to those available from our existing argon-ion sources, allowing investigation of different materials and the effects of using different wavelengths for the address and hold functions.

4.3 Staff Support

Funds in this heading have been used to support one full time senior staff member; in addition several other staff including senior and junior University supported research and academic staff and graduate students have actively contributed to the project.

4.4 Other Costs

Additional funds (as specified in the original proposal) have been used to supply prototype zinc selenide optical logic elements to the Dayton Research Institute and to the Naval Ocean Systems Centre in San Diego.

References

[1] Coupled Wave Theory for Multiply Exposed Thick Holographic Gratings, S.K. Case, JOSA 65 p724 1975.

[2] Coupling in Doubly Exposed Thick Holographic Gratings, R.A. Alferness and S.K. Case, JOSA 65 p730 1975.

RECENT PUBLICATIONS

1. S.D. Smith, J.G.H. Mathew, M.R. Taghizadeh, F.A.P. Tooley and A.C. Walker
   Demonstration of a Triple Bistable Element Loop Circuit for a Digital Parallel All-Optical Computer
   Optical Bistability III, Proceedings of the Topical Meeting, Tucson, Arizona, Dec. 1985, Editors: H.M. Gibbs, P. Mandel, N. Peyghambarian and S.D. Smith (Springer-Verlag, 1986), p. 8.

2. B.S. Wherrett
   All-Optical Computing: Circuit and Component designs
   ibid, p. 12.

3. A.K. Kar and B.S. Wherrett
   Bulk ZnSe: Linear Transmission to Damage through Dispersive Bistability and Absorptive Switching
   ibid, p. 136.

4. A.K. Kar, H.A. MacKenzie, Wei Ji, J.J.E. Reid, R. Grisar, D. Ball, H.M. Preier
   Optical Bistability in PbSnSe at Room Temperature with Infrared Radiation at Milliwatt Powers
   ibid, p. 144.

5. Wei Ji, A.K. Kar, U. Keller, J.G.H. Mathew and A.C. Walker
   Cascaded Bistable Optical Devices Based on Two-Photon Absorption in Room-Temperature InSb
   ibid, p. 35.

6. M.R. Taghizadeh, F.A.P. Tooley, J.G.H. Mathew
   White Light Switching of Visible Nonlinear Interference Filters and its Implications for the Design of Spatial Light Modulators
   ibid, p. 45.

7. H.A. Al-Attar, W.J. Firth, H.A. MacKenzie, F.A.P. Tooley and A.C. Walker
   Gain Bandwidth Product of an InSb Transphasor
   ibid, p. 49.

8. J.G.H. Mathew, M.R. Taghizadeh, E. Abraham, I. Janossy and S.D. Smith
   Observation and Analysis of Critical Slowing Down in Nonlinear Visible Interference Filters
   ibid, p. 57.

9. D.J. Hagan, I. Galbraith, H.A. MacKenzie, W.J. Firth, A.C. Walker, J. Young, J.J.E. Reid and S.D. Smith
   Measurement of Transverse Coupling between adjacent InSb Optical Switching Elements
   ibid, p. 189.

10. W.J. Firth, I. Galbraith and E.M. Wright
Diffusion Effects in Optically Bistable Arrays
ibid, p. 193.

11. A.C. Walker, S. Aitchison, J.T. Chilwell, S.T.D. Ritchie, P.M. Rodgers
Intrinsic Optical Bistability in Passive GaAlAs Waveguide
ibid

12. I. Janossy, M.R. Taghizadeh and E. Abraham
Laser-induced Distortion of Nematic Liquid Crystal Films and Observation of Cavityless Optical Bistability due to Thermal Effects
ibid, p. 160.

13. J.V. Moloney, J. Ariyasu, C.T. Seaton and G.I. Stegeman
Stability of Nonlinear Stationary Waves Guided by a Thin-Film Bounded by Nonlinear Media
Applied Physics Letters (in press).

14. J.V. Moloney, J. Ariyasu, C.T. Seaton and G.I. Stegeman
Numerical Evidence for Non-Stationary Nonlinear Guided Waves
Optics Letters (in press).

15. J. Ariyasu, C.T. Seaton, G.I. Stegeman and J.V. Moloney
New Theoretical Developments in Nonlinear Guided Waves: Stability of TE_1 Branches
J. Quant. Electron., June 1986 (in press).

16. B.S. Wherrett, D. Hutchings and D. Russell
Optically Bistable Interference Filters:- Optimisation Considerations
J. Opt. Soc. Am., 3, 351-362 (1986).

17. A.K. Kar and B.S. Wherrett
Thermal Dispersive Optical Bistability and Absorptive Bistability in Bulk ZnSe
J. Opt. Soc. Am., 3, 345-350 (1986).

18. B.S. Wherrett, A.K. Kar, D. Russell, D. Hutchings and H. Clement
Optimisation of Optically Bistable Interference Filters for Array Processing
Optica Acta, 33, 517 (1986).

19. A.C. Walker
Application of Bistable Optical Logic Gate Arrays to all-Optical Digital Parallel Processing
Appl. Optics., 59 (1986).

20. S.D. Smith, A.C. Walker, B.S. Wherrett, F.A.P. Tooley, J.G.H. Mathew, M.R. Taghizadeh and I. Janossy
Cascadable Digital Optical Logic Circuit Elements in the Visible and Infrared: Demonstration of Some First All-Optical Circuits
Appl. Optics, 59 (1986).

1

INVESTIGATION AND DEVELOPMENT OF ARCHITECTURES AND INTERFACING ELEMENTS APPLICABLE TO ULTRA HIGH SPEED OPTICAL COMPUTER SYSTEMS

Report 6371, Issue 1, May 86

Contract No. N00014-85-K-0479

FERRANTI
Computer Systems

FERRANTI COMPUTER SYSTEMS LIMITED
Bracknell Division, Western Rd., Bracknell, Berks. RG12 1RA
Telephone: Bracknell (0344) 483232 Telex: 848117

INVESTIGATION & DEVELOPMENT OF ARCHITECTURES AND INTERFACING ELEMENTS APPLICABLE TO ULTRA HIGH SPEED OPTICAL COMPUTING

ABSTRACT

This report identifies important criteria for the design, development and implementation of complex computer systems. Three candidate architecture types, described as Systolic, Data Flow and Massively Parallel Architectures are appraised and their relevance and application discussed.

An assessment and evaluation of technologies suitable for discrete optical computing is described, where particular emphasis has applied to ZnSe non-linear and bistable interferometers. Engineering specifications for these, together with supporting technologies, have been derived. The specifications indicate target performance requirements for the components and interconnection devices identified for implementing parallel architectures.

An analysis of cascaded interferometers, configured to illustrate the minimum conditions for general purpose applications, has resulted in the derivation of general expressions which describe minimum acceptable levels for the transfer characteristics of these devices. These expressions also portray trade off mechanisms which can be exploited to achieve improved performance. Recommendations are made for improved performance of ZnSe interferometers.

A design study for implementing the Berlekamp-Massey Algorithm using ZnSe interferometer technology is described, resulting in an extension of a systolic architecture. This may successfully exploit the potential parallelism within each etalon to decode very long sequences of binary-valued syndromes.

CONTENTS LIST

SECTION A: PHASE 1 FINAL REPORT

SECTION B: RECOMMENDATIONS FOR FUTURE PROGRAMS

SECTION A: PHASE 1 FINAL REPORT

1. INTRODUCTION

Section A of this report details the principal activities and conclusions of a research and development program carried out under contract to the University of Dayton, Ohio, in support of the Strategic Defence Initiative administered by the U.S. Office of Naval Research, with additional capital and resources provided by Ferranti Computer Systems Limited.

The initial phase of the research program considers the essential attributes of more general purpose processing functions suitable for high performance real time applications, and concentrates on suitable device implementation based upon non-linear interferometer techniques and associated technologies.

Acknowledgement is given to the research team at Heriot-Watt University for their specialist advice and consultation concerned with ZnSe optically bistable interferometers.

Section 5 extracts the key areas of importance and reviews the achievements resulting from the initial phase of the research program. Sections 2, 3 and 4 describe the work in detail.

Under Section 2 of this report applications and architectures, which could exploit large parallelism and dense global interconnections, are discussed. This addresses prominent areas resulting from an extensive review of current and potential techniques and philosophy, and naturally benefits from the extensive experience obtained from related activities within the Company.

Under Section 3 optical and electro-optical technologies are discussed, with particular emphasis towards the utilisation of non-linear interferometers. The major part of this work has been to interpret the requirements of the processing functions under study and equate to the potential performance of the technologies. Practical considerations for final implementations have significantly influenced the conclusions.

Section 4 details a case study development which implements the Berlekamp-Massey Algorithm using etalon-based optical technology.

The remainder of Section 1 outlines the statement of work, and the approach and organisation of the research effort.

Future research topics are described under Section B of this report.

1.1 STATEMENT OF WORK

We anticipate that the long term objectives of the research program are to establish a processing capability, produced using optical components, which perform computation at rates many orders of magnitude faster than contemporary electronic systems.

An effective program relies on extensive interaction to equate the device characteristics to the various system requirements.

Heriot-Watt University has declared that it would welcome close collaboration with Ferranti Computer Systems Limited, to focus their research towards specific system areas and provide additional resources to develop techniques for applications.

Ferranti Computer Systems Limited proposes a schedule of work which, from the onset, considers the implications of environment, interfacing, control elemental and full system structuring, as appropriate to the

processing requirement of a real-time system. This approach will highlight specific objectives, and qualify the relative merits of different research program options. The process is iterative as advances in technology and concepts emerge.

The first phase of the schedule is detailed below and covers the period 15th July to 31st December 1985.

<u>Phase 1 - Data and Control Interface Study for an Array of Processing Elements</u>

The objective of the initial study is to identify those performance characteristics which qualify or constrain the techniques which can be employed in implementing required system functions.

Particular emphasis will be placed on methods for entering data and control to an array system, as successful implementation of these interfaces will greatly enhance the progress in subsequent research activities. The study will consider coherent and non-coherent optical sources and electro-optical interfaces where they are considered relevant to this phase.

The study, and subsequent work, will be based upon the research activities conducted at Heriot-Watt University.

It is essential that Ferranti Computer Systems Limited performs experiments which complement the research at Heriot-Watt, if satisfactory progress is to be maintained. It is proposed, therefore, to establish a suitable trials system operating with visible light, on the Ferranti Computer System Limited site.

The objectives of Phase 1 are to identify those techniques suitable for interfacing data and control to an array of elements and to make proposals for the work content of Phase 2.

1.2 RESEARCH PHILOSOPHY AND ORGANISATION

1.2.1 Research Philosophy

The principal objectives of the research program are to exploit the emerging technologies and establish innovative approaches which will lead to optical computer systems which can process at orders of magnitude greater than contemporary electronic systems. It is envisaged that future systems will be able to harness the unique qualities of optical components in areas where electronic techniques are already meeting practical and fundamental limitations.

It is already recognised that to simply replace existing electronic structures and architectures with ultra-high speed optical equivalents is unlikely to achieve the required improvements. It is clear that for maximum advantage new structures, architectures, algorithms, manufacturing processes and development methodologies must be pursued.

Any computer system is extremely complex. Integrated circuit technology may be identified as one of the most important developments to enable rapid evolvement of electronic computer techniques, but of equal importance was the development of supporting technologies, standardisation, versatile devices and configurations, and automatic tools for simulation, modelling and programming - all of which provided an acceptable framework so that tradeoffs can be perceived and optimisation achieved in a rapid and efficient manner.

Experience has shown that new technologies and techniques can only be successfully harnessed if positive and negative tradeoffs are understood and the resulting performance is optimised for maximum benefits. The most effective program of research relies on clearly defined targets, based upon relevant applications, which take account of those aspects together with a broad perspective of possible approaches, which may be

developed to overcome specific practical shortcomings, and hence achieve optimisation. An applications driven philosophy has therefore been adopted for the research program.

1.2.2 Organisation

The Company is established to research and develop components, products, systems and software for civil and military applications, including computers and peripheral equipment. This work is carried out in a multi-disciplined environment where each area employs experts and specialists employed to advance techniques and capability.

For the initial phase, the research program had brought together specialists in advanced techniques relating to electronic and optical technologies, devices and their design, and to computer applications, architectures and software. In addition the program has gained benefit from extensive practical and intellectual interaction from concurrent research activities within the company.

Fluid communication has been maintained to derive a common understanding for the requirements under consideration and for the practical and potential performance factors that will influence successful implementation.

2. APPLICATIONS AND ARCHITECTURES

2.1 SCOPE AND OBJECTIVES

A major objective in this investigation was the rapid identification of those architectures and applications which might:

(a) appear to benefit from the advantages offered by optical processing (such as large parallelism and dense global interconnections),

and (b) show reasonable potential for implementation in some combination of optical or opto-electronic techniques,

and (c) appear useful in the context of existing or forseeable systems.

These are generally subjective criteria, and hence many of the decisions made in the selection process are necessarily based upon experience and intuitive judgements. However, it is hoped that our conclusions and suggestions might usefully influence and support the direction of future research into optical computing structures.

In addition to our more general objectives we undertook to respond to a specific suggestion by the University of Dayton that we consider the suitability of the Berlekamp-Massey algorithm, as an application which might benefit from a design implemented using an etalon-based optical technology, such as that under development at Heriot-Watt University in Edinburgh.

In considering applications and architectures, it is important to maintain a perspective on the system as a whole. Hence, in the remainder of section 2, we commence by discussing some general system requirements, which are of particular relevance to real-time defense systems

such as those designed and built here at Ferranti Computer Systems Ltd. We then go on to describe some of the more important trends in parallel computing and architectures, and the significance of these in an overall system context. Finally, we propose some candidate architectures which we believe meet the criteria outlined above.

2.2 GENERAL SYSTEMS REQUIREMENTS

2.2.1 Introduction

The concept of the Strategic Defense Initiative is probably the most demanding challenge to computer systems technology yet made. The scale and complexity of the proposed system is of unprecedented proportion, and yet the strategic significance of the system necessitates that the very highest levels of secure and dependable operation must be attained.

In order that the development of the SDI system can be managed and controlled to meet these objectives on time, within cost, and to the rigorous satisfaction of the systems requirements, a rapid development in current methods of system design is essential. It cannot be overstressed that many requirements of the SDI programme, e.g. multi-level security, system reliability, fault-tolerant operation, etc. can not be achieved by simply adding on functionality to an existing or future system which does not possess these attributes. The required functionality must be designed into the system from the very beginning. For example, the integrity of a system (i.e. the correctness of the data which the system outputs) encompasses the <u>total</u> system design and is only achievable by detailed consideration of the aspects of error detection, error containment, error reporting, fault reporting and error recovery. Error recovery may involve an algorithmic or hardware reconfiguration. Similarly, multi-level security of a system requires the segregation of data areas. Both the examples above are directly affected by, for example the choice of system architecture. Architectural considerations are discussed further in section 2.3.

2.2.2 Modularity

Modularity theory has been successfully employed to reduce and control the complexity of both hardware and software aspects of systems, with consequent benefits in achieving required degrees of maintainability, reliability, etc. when used together with an appropriate methodology.

The decomposition of a system into logical and physical modules has traditionally been aimed at reducing system complexity in order to aid the management and control of system development. Such a decomposition is often performed according to identifiable functional components, that is, a system requirement specification is functionally analysed and functionally decomposed into hardware and software modules. With increasing demands upon the requirements of modern computer systems, modularization techniques have proceeded to further dimensions. One aspect of this is discussed later in relation to reliability, recovery and reconfiguration.

The correct choice of modular decomposition is of fundamental importance when addressing system requirements such as formal verification methods, security, fault tolerance, maintainability, and so on.

2.2.3 Standardization

The importance of standardization in the engineering of systems is now widely recognised. A number of national and international committees have been established in recent years, to determine appropriate standards to be applied in particular areas. For example, the programming language ADA has been adopted by the U.S. Department of Defense as the standard language for software production in military systems. Another example is the I.S.O. 7-layer model of the communications interface in computer systems.

Reasons for standardization include compatibility between different vendors of system components, multi-sourcing of components to improve availability, and to provide a common framework for the discussion, analysis, design, interfacing and specification of systems. In addition, suitable standards can result in the detailed specification of a product, in terms of function, interfaces, performance and operating characteristics.

On a smaller scale, the members of a programming team obviously need to agree standards of communication and documentation between themselves, and when supplying the completed system to the user. Significant advantages are gained by an organization having standards and standard procedures, e.g. reduction in the need for staff re-training, staff mobility, and experience sharing. Also, the enforcement of common documentation standards simplifies the handing over of the product. The standard can be drawn up to suit procurement methods and can be officially controlled.

2.2.4 Fault Tolerance

Fault tolerance is concerned basically with the issues of reliability, recovery and reconfiguration. These are by no means independent issues, and the three topics are difficult to discuss in isolation.

Randell, Lee and Treleaven [RAND 78] have surveyed the issues involved in achieving high reliability from complex computing systems, and discuss the relationship between system structuring techniques and techniques of fault tolerance. The topics they cover include:

(a) redundancy in hardware and software

(b) the use of atomic actions to structure system activity so as to limit information flow and provide recovery points

(c) strategies for the location and handling of faults and for damage assessment , and

(d) forward and backward error recovery techniques.

The definitions which Randell, Lee and Treleaven have proposed help to distinguish between reliability and availability, and between failures, errors and faults. They state that:-

"The RELIABILITY of a system is taken to be a measure of the success with which the system conforms to some authoritative specification of its behaviour. Without such specification, nothing can be said about the reliability of the system. When the behaviour deviates from that which is specified for it, this is called a FAILURE. A failure is thus an event, with the reliability of the system being inversely related to the frequency of such events. Various formal measures related to a system's reliability can be based on the actual (or predicted) incidence of failures, and their consequences. These measures include Mean Time Between Failures (MTBF), Mean Time To Repair (MTTR), and AVAILABILITY, that is the fraction of the time that a system meets its specification. We term an internal state of a system an ERRONEOUS STATE when the state is such that there exists circumstances (within the specification of the use of the system) in which further processing, by the normal algorithms of the system, will lead to a failure which we do not attribute to a subsequent fault. The term ERROR is used to designate that part of the state which is incorrect. A FAULT is the mechanical or algorithmic cause of an error".

The authors point out that it can be extremely difficult to attribute a given failure to a specific fault, since a detected error is merely a symptom of the fault which caused it. This is particularly true of software faults which stem from unmastered complexity in the system design. Reliability of a physical system can never be formally proved, and a perfectly reliable system can never be achieved. The need therefore arises to develop hardware and software techniques for improving reliability.

System failure is defined as a deviation from the requirements specification. If system reliability is to be precisely defined, the system specification must be complete and unambiguous (i.e. well defined). A system specification would only very rarely be well defined. In most cases, it would be impossible to specify the exact correspondence between system outputs and inputs. Consequently, it would not normally be possible to deduce from the system specification whether a particular output value is correct response to a given value of the input. The detection of a system failure would therefore be limited somewhat to the more obvious deviations from the system specifications, i.e. deviations such as:-

(a) An incomplete output,

(b) an output is outside a specified range,

(c) an output is not obtained within a specified time limit, or an output is completely absent. etc.

System failure occurs when the following three factors coincide:-

(a) The system contains a fault.

(b) Data enters the system, and when processed by the faulty part of the system, generates a fault.

(c) The system does not contain an error recovery algorithm which is able to cope with the particular error.

It is impossible to predict when these three factors will coincide , thus system failure would be expected to be a random event.

It may happen that inputs may exist which are outside specified ranges. A system should then not only be reliable (perform according to its specification, but should also be ROBUST, ie. it is able to process without error some input data outside its specification. This requirement of robustness is important in systems used in an interactive mode when users may utilise the system in a manner not anticipated by the system designer. Robustness is also important in real-time systems when spurious inputs are unexpectedly generated.

It must be accepted that absolute correctness of a complex software system cannot be attained. Even if the code produced could be guaranteed correct, faults can still effectively arise in it as the result of permanent hardware malfunctions, glitches, power surges, temporary hardware malfunctions due to e.g. electromagnetic or nuclear irradiations, transients, user errors, etc. To ensure total system reliability, fault tolerance must therefore be designed into all aspects of the system, hardware, firmware and software. It should be noted that fault tolerance is not achieved without overheads of space and time in the run-time system. The various stages of fault tolerance, detection of an error, recovery (and perhaps even reconfiguration of the system) will be relatively time consuming and detrimental to the performance of a real-time system. As with all systems, a compromise will be required between the degree of total system fault tolerance which is provided and the consequent run-time overhead introduced.

Some examples of fault tolerant techniques are the use of recovery blocks, process checkpointing, rollback and recovery, and the use of redundant information to detect and correct errors. Redundancy is the key to achieving fault tolerant software, but redundancy of itself will be useless unless it can be exploited at an acceptable cost in terms of extra storage, execution times and I/O operations. The proper deployment of redundancy allows errors to be detected in the data structures of a system.

2.2.5 Availability

Availability is a function of the network architecture in the sense that multiple physical nodes are required to achieve very high availability. A failure must be confined to that physical node in which the failure has occurred. If a system can be readily reconfigured to isolate a faulty node, then in a system which contains a surfeit of the nodes required to process a given application, it would almost certainly not be cost effective to produce nodes of high availability. The system would be required to reconfigure itself and continue processing after a node failure. Probably more difficult than reconfiguring the degraded system would be the task of incorporating a repaired node back into the system. The main difficulty would be the reconstruction of the stable data base without causing great perturbations to the normal system operation.

Reliability is a parameter of the physical node. Any extra facilities provided in a node to enhance reliability must be cost effective in terms of reducing maintenance costs.

The integrity of a system, i.e. the correctness of the data which the system outputs, encompasses the total system design. Integrity is only achieved by detailed consideration of all the following topics:-

(a) Error detection. If errors cannot be detected then incorrect results may be produced. Error detection is the most imporotant consideration in system design since there is little point in having sophisticated error recovery mechanisms, reconfiguration schemes, error reporting facilities, etc unless all errors are first of all detected.

(b) Error containment. When errors do occur, it is desirable to prevent them propagating so that minimum corruption of the system occurs. Errors should be confined to, at the least, a physical node. If errors are kept confined to lower levels, then two benefits are

obtained. The first benefit arises in that if one error is prevented from causing further errors, then confinement within a node is made easier. The second benefit is that the ease of diagnosing a fault is enhanced.

(c) Error reporting. Errors need to be reported as quickly and as accurately as possible. It is essential that the knowledge that data may be incorrect be known, even if it cannot be automatically corrected.

(d) Fault reporting. Not all faults produce errors (e.g. corrective action may be taken). Thus, while fault reporting is related to error reporting, it is not synonymous with it. Fault reporting is valuable in warning of potential future failures.

(e) Error recovery. So that a system can continue functioning after an error has been detected, a method of recovering from the error must be available. Recovery is often achieved by replicating data contained in one node into another node, or by maintaining a history of data from which a correct version can be derived. A single system could use both methods.

Even a brief survey of the literature regarding fault tolerant systems reveals a heavy bias towards hardware faults. The obvious fact that algorithmic faults (be they in hardware or software) are of relatively monumental importance is all too often ignored. It is relatively easy to ensure that a hardware fault will not affect two physical nodes simultaneously, but an algorithmic fault can very easily affect many nodes. It should be borne in mind that the failure to detect an error is always an algorithmic fault.

The efficient design and implementation of a high availability system ideally requires that the language used supports the requirements of a high availability system. For example, Ada has Exception Handling to

provide the means to contain and report detected faults and errors in a well structured manner. However, while it provides assistance in preventing errors by detecting those faults which may cause them, no great assistance is given in detecting faults in the logical part of the system or in the software design. An assessment of languages is needed to discover what support can be given to the fault tolerant system designer.

2.2.6 Correctness

Many faults arise from human errors made during the formulation of requirements specifications and the early design stages. It is therefore a benefit to thoroughly check for correctness at every stage of system and software development. The systematic application of fault avoidance techniques, when supported by careful checking for correctness of each design step, can reduce the number of residual system/software faults due to human error down to neglible proportions. The cost of thorough checking for correctness at every development stage would be expected to be considerably less than the costs due to subsequent debugging.

In principle, it is possible to formally verify the correctness of software, but in practice many difficulties exist, despite much research effort. It is possible to verify correctness because, software, being a system of statements expressed in a formal language, can use the formal techniques of mathematical logic to prove the specification is completely met. The processes are similar to that used in proving an algebraic theorem. Among the daunting practical difficulties encountered are:-

(a) Formal specification languages are not yet fully developed.

(b) The description of a user's requirement by a formal language proves to be difficult in practice.

(c) The methods used for correctness proving are long and tedious, and would benefit from automation.

2.3 PARALLEL COMPUTING AND APPLICATIONS

It is now widely believed that the sequential von Neumann model of computation is likely to be inadequate to meet the demanding requirements of many future systems [TREL84]. In the design of current systems there is a growing emphasis on the use of distributed and parallel processing for reasons which include lower costs, improved performance, and fault-tolerance.

Over the past few decades, considerable research effort has been focussed upon the problems associated with parallel computing, for which the main issues now seem to be: choice of machine organization, choice of computational model, and choice of programming style. As a result many architectural proposals have been made, some of which have led to a concentration of activity in specific areas.

Some particular problem classes have been approached through the development of architectures "tailored-to-fit" the problem. Most notable of these perhaps is the systolic array philosophy, in which the design limitations of VLSI are exploited in the hardware implementation of algorithms. Usually, such algorithms exhibit a highly regular structure as typified by linear algebra problems, and are generally "computationally intensive" (see section 2.4.1).

The so-called "supercomputers" (e.g. CRAY-1, CDC6600, etc. [NORR84]), besides possessing the full flexibility of a conventional von Neumann style architecture, generally employ a number of independent arithmetic units in a pipeline arrangement. This can provide a very high performance in the execution of certain regular numerical algorithms such as the numerical solution of partial differential equations (e.g. in weather prediction), however achieving such performance can prove very difficult in practice and often requires the programmer to have an intimate knowledge of the machine architecture.

Parallel architectures falling into the MIMD (Multiple Instruction, Multiple Data) category usually support a greater degree of concurrency than the more conventional supercomputers described above. A recent survey of MIMD computers in the United States by Hockney !HOCK84§ defines MIMD computers as: "control-flow computers capable of processing more than one stream of instructions" (note that this definition excludes data-driven and demand-driven architectures !TREL82§). Hockney's paper goes on to propose a structural classification of MIMD computers, in which a distinction is made between networks and switched systems, shared memory and distributed memory systems, bus and cross-bar and multi-stage switched systems etc. Hockney observes that the Multi-stage, shared-memory, switched computer is currently the most favoured MIMD architecture (e.g. TRAC - Texas Reconfigurable Array Computer).

The major disadvantage with the architectures discussed so far is that they can only exploit explicit parallelism, that is, when it is specified explicitly in the program. Such a restriction is tolerable only in simple cases where the parallelism in an algorithm is easily determined and synchronisation is correctly specified. Furthermore, the so-called "software crisis" (i.e. the escalating productivity problems associated with software, due to system complexity and shortage of skilled staff) clearly demonstrates that an extra dimension of complexity in the explicit specification of parallelism is the last thing we need in software engineering. Of course this argument is especially applicable to the case of large and complex systems such as that proposed in the U.S. Strategic Defense Initiative !PARN85§. There are indications however, that the "explicit" approach may become feasible through the use of languages such as OCCAM !INMO84§, which has a formal mathematical foundation, and is based on Hoare's CSP!HOAR78§. These languages are founded on a model of computation which is essentially a generalisation of the von Neumann model, embracing such concepts as decentralization of control !TREL83§, explicit parallelism, and communicating processes.

A slightly different approach to solving the above problem is found in the Functional style of programming, as argued in Backus' classic paper [BACK78]. A program written in a Functional language (e.g. FP, "pure LISP", HOPE) contains implicit parallelism, and can be regarded as a nested expression containing functions applied to their arguments. Functional programs possess some important mathematical properties which are generally absent from conventional procedural languages [TREL86]. For instance a program and the result of executing that program are mathematically equivalent, and may be regarded as the same object expressed in different forms. Another property is "referential transparency", which means that the value of an expression is determined by its definition only, and not by previous invocations of that expression or any other expression. Finally, the absence of "side-effects" in a Functional program means that such concepts as the assignment statement (perhaps the single most important "bottleneck" in the von Neumann procedural programming model) are not permitted. These properties result in a style of programming having a sound mathematical basis in which the notion of "executable specifications", through the use of techniques such as program transformation, becomes a very real possibility.

Architectural support for the Functional style of programming is provided by parallel machines employing both data-driven and demand-driven models of computation [TREL82]. In general, data flow computers employ a data-driven model of computation and reduction computers employ a demand-driven model of computation. It is worth noting perhaps, that the Japanese Fifth Generation project is developing both data flow and reduction machines (PIM-R and PIM-D [TREL86]) at the ICOT laboratories. However, the most extensive research into data flow and reduction machines has taken place at MIT (data flow), University of Manchester (data flow), and Imperial College London (reduction). Such architectures are capable of supporting those general applications which have previously been implemented in conventional single processor (von Neumann) architectures, in addition to the many Artificial Intelligence applications currently emerging.

A novel direction in machines for the study of natural and artificial intelligence is found in massively parallel architectures and the "connectionist" model of computation. There are several different approaches in this area, but the general concepts shared by most seem to include: massive parallelism with relatively slow processing nodes; emphasis on the establishing, strength, and reciprocal feedback of interconnections; behaviour through evolution into some stable state (rather than following a programmed sequence). The Hopfield model [HOPF82] is a notable and influential example, and the Boltzmann Machine concept proposed by Hinton et al [ACKL85] is currently receiving considerable attention. In the field of Artifical Intelligence, an important area of concern is the way in which "real-world knowledge" should be represented in a machine. A recent approach which is now receiving much attention, is the use of a scheme termed "semantic networks". In such a network nodes (vertices) represent concepts, and the links (edges) between the nodes represent relationships between concepts. At Carnegie-Mellon University, Fahlman proposed a system called NETL [FAHL79] in which a semantic network is represented directly in parallel hardware, rather than a software simulation as would be necessary in a conventional von Neumann architecture. Such ideas have motivated the development of the "Connection Machine" by Hillis and others at MIT [HILL85]. Interestingly, this machine can be programmed in a variant of LISP and appears to be useful in a range of applications, in addition to the direct implementation of semantic networks.

In summary, we should first note that in general, no one architecture is optimally suited to all applications. Some architectures are designed for the solution of a very specific class of problems, whereas others are designed to perform quite well over a range of applications but are not really outstanding in any particular area. For example, the CRAY-1 "supercomputer" is extremely powerful when performing floating-point arithmetic operations on long vector operands, but is relatively poor in terms of symbolic processing capability. The very popular LISP (list-processing) language requires such a capability which typically

involves the dynamic allocation and release of storage ("garbage collection"), and data structure manipulation and list processing activities (e.g. multiple consecutive memory accesses to follow a chain to pointers or addresses). The SYMBOLICS family of computers were developed from work originally carried out at MIT, with the specific aim of supporting this type of processing application [TREL86]. These two examples represent extreme cases and in practice, most systems will require an appropriate balance of capabilities which are suited to the application in hand. However, achieving a suitable balance is often very difficult, and must be based upon assumptions concerning possible/likely events in the real world and the system, and how the system is required to respond to these events.

Concerning architectural styles, we may draw the following general conclusions:

Architectural Style	Applications
Systolic Arrays	Fixed, regular algorithms both numerical and non-numerical. Very high performance but inflexible.
Supercomputers	Support general applications with average performance and certain numerical applications with very high performance.
MIMD computers	Support general applications but performance frequently very dependent on particular architecture and algorithm used. Potentially very high performance but often difficult to achieve.
Data Flow and Reduction computers	Support general applications and seem well suited to symbolic processing. Performance potential is very high due to "fine grain" parallelism, but is still to be fully demonstrated.
Massively Parallel computers	Support general applications to a limited extent but are particularly aimed at Artificial Intelligence applications. Potentially very high performance but real applications examples need to be demonstrated.

In considering the role of optics in the above discussions, it seems reasonable to conclude that the same fundamental computing issues are still applicable. That is, we must still consider the machine organization, computational model, and programming style. In addition, there are questions concerning number representation, control techniques, arithmetic techniques, and optical "circuit" techniques (i.e. low-level building blocks) which must also be resolved. Optimistically, we may expect to gain a tremendous advantage over electronic systems in particular areas such as massive parallelism, dense global interconnections and parallel associative memory systems.

2.4 SOME CANDIDATE ARCHITECTURES

At the hardware organisation level, an optical digital computer system will radically differ from an electronic system. The optical computer system will not be realized by merely substituting functionally equivalent optical devices for electronic devices. New criteria are needed to establish the optimal trade-off between gate count and complexity of interconnection. These different design criteria must frequently suggest new physical architectures which exploit the advantages of an optical implementation, and perhaps where electronic implementations would prove impractical. For instance, architectures which are communication intensive may often transfer easily onto relatively simple optical systems, even though communication limitations are already imposing severe restrictions upon VLSI architectures.

Optical techniques offer the possibility of designing parallel architectures with a number of mechanisms for implementing interconnections and communications. However, the realization of this potential requires the accelerated development of optical devices, parallel algorithms and appropriate architectures.

This section briefly describes three general architecture styles which we believe might prove suitable for an optical implementation, and as such warrant further investigation.

2.4.1 Systolic Architectures

The term "systolic" is borrowed from physiology. Just as blood is pulsed around the circulatory system, so in a systolic array data is made to flow in a regular, pulsed way through an array of processing elements. The systolic array is a hardware realization of an algorithm, employing a high degree of pipelining together with parallel processing, which achieves a very high computational throughput [KUNG79].

In general, systolic arrays have the following characteristics:

(a) They are implemented with a few types of simple cells.

(b) Data and control flow through the array is simple and regular.

(c) The array makes use of extensive pipelining and concurrency as several data streams move through fixed paths.

(d) The array maximizes the use of each input data item, and a high computation throughput is achieved.

"Data intensive" computations are not necessarily suitable for use with a systolic array, since contention would arise with a large number of data items flowing around a relatively small (limited by the number of computational operations) array of processing elements.

On the other hand, "computation intensive" computations are very suitable for processing with an array designed for the particular algorithm. An example of a computation intensive problem is found in matrix multiplication. It is computation which predominates I/O in matrix multiplication, and takes on an increasing proportion of the total operations as the matrices grow larger.

In a general purpose matrix multiplication, of say a (p x q) matrix A by a (q x r) matrix B to give a (p x r) matrix C, then since each element in C is given by

$$C_{ik} = \sum_{j=1}^{q} a_{ij} \, b_{jk}$$

each processing element is required to:

(a) input a pair of numbers, a_{ij} and b_{jk},

(b) perform a multiply to obtain a_{ij} b_{jk},

(c) add the result into an accumulator (initialized to zero),

(d) repeating steps (a), (b) and (c) q times in all,

(e) output the final contents of the accumulator, and

(f) reset the accumulator to zero.

The systolic array required to perform the matrix multiplication would consist of an array of (p x r) such processing elements. There would be totals of (pq + qr + pr) I/O operations, (qpr) adding-to-accumulator operations and qpr multiplications. The multiplication of a single pair of matrices would require a total of (pr) hardware pulses, but if a stream of matrix pairs were multiplied, throughput would increase by a factor of q giving effectively one matrix multiplication for every (pr/q) pulses.

If we were dealing with square matrices, say of order (pxp), then the ratio of computational operations to I/O operations is (2p/3), so that the computation is indeed "computation intensive" and increasingly so with increasing (p). The (p x p) systolic array would complete a single

matrix multiplication is (p^2) hardware pulses, and for a stream of multiplications, would provide (p) results in this same time giving an effective figure of one multiplication per (p) pulses. Execution time thus increases linearly with the order of the square matrix, a very different case to sequential execution on a single processor.

The systolic approach to the solution of many demanding real-time applications problems has been studied by some military and industrial departments in the USA. Both the Naval Ocean Systems Centre (NOSC) at San Diego, California, and ESL Inc of Sunnyvale, California have found the systolic approach superior to many other parallel or pipelined processors for certain applications.

NOSC has built a hardware version of a systolic array processor comprising a dynamically reconfigurable array of 64 processing elements, each of a single board. Each element is a true independent processor capable of 32 bit floating point arithmetic. The array is intended for use as a test bed for the evaluation of array configurations and algorithms for performing different tasks. ESL has produced a 28 element systolic processor for evaluation purposes.

The development of the systolic approach has been rapid since the initial work by Kung at Carnegie-Mellon Univerity in 1978. Kung has since designed a dynamically programmable VLSI systolic chip, and research is now proceeding along two fronts - hardware and software. Software systems are being developed to form the data paths within the chip, i.e. to produce the array configuration required for the application. ESL has developed a compiler to perform this task for their system.

We believe that systolic architectures could be developed for optical or hybrid implementation in timescales which are relatively short compared with data flow (section 2.4.2) and massively parallel architectures (section 2.4.3). This is primarily because of the degree of flexibility inherent in the respective approaches - the systolic architecture

clearly being the least flexible. It is not immediately apparent that global interconnection capability is necessary in a systolic architecture. On the contrary, one of the motivations behind a systolic approach is the very fact that global interconnection is difficult and expensive in a VLSI context. However, we take the view that if such a capability is available (e.g. in optics) then we should expoit this feature also, extending the power and applicability of the systolic concept. An example of this is seen in section 4 where we discuss a systolic optical implementation of the Berlekamp-Massey algorithm. Our implementation exploits the availability of global interconnections to provide a (potentially) large array of systolic cells but using a configuration for only a single cell, different regions of the focal plane functioning as different cells. Although we demonstrate the technique only for a one-dimensional array (vector) of systolic cells it is not difficult to see how it could be extended to the case of two dimensional arrays. We also see potential for the introduction of hardware redundancy using this method, leading to a fault-tolerant capability not normally associated with systolic architectures implemented using VLSI technology.

Systolic VLSI designs using one or two dimensional arrays or tree structures have been implemented, and shown to perform well in the following fields:-

(a) Signal and Image Processing, for

(i) FIR and IIR filtering and one dimensional convolution.

(ii) Discrete Fourier Transforms

(iii) One dimensional and two dimensional filtering.

(b) Matrix Arithmetic, for

(i) Matrix-vector and matrix-matrix multiplication.

(ii) Matrix triangularization (solution of linear systems, matrix inversion).

(iii) QR decomposition (Eigenvalues, least squares computations).

(c) Non-Numeric Applications, such as

(i) Language recognition - string matching and regular expressions.

(ii) Polynomial algorithms - polynomial multiplication and division and polynomial greatest common divisor.

(iii) Relational data base operations.

2.4.2 Data Flow Architectures

During the last decade, a number of novel computer architectures based on new intrinsically parallel models of computation have been proposed [TREL82], some of which have actually been constructed. The main stimulus for this work has come from Dennis' data flow concepts [DENN74], Backus [BACK73] and Berkling's Functional languages and machines [BERK75]. The resulting architectures can be broadly classified as either data-driven or demand-driven. This discussion of data flow architectures is based upon the description in [TREL82].

Data flow represents a radical departure from the von Neumann approach to computer organization. Rather than a centralized control unit containing a "program counter" register which controls the sequencing of instructions, the selection of instructions for execution is determined

only by the flow of data between instructions. In this way, the control is distributed and many operations may be executed in parallel. The data flow approach has several attractive implications for computer architecture because of its simple operational semantics and potential for parallelism.

Data flow computers occur in many forms. However, basic to all data flow machines is a mechanism by which instructions are executed immediately all their required operands are available. In the logical system, we may regard each instruction as having continuous access to a processing element, simply waiting for operands to arrive. The key factor governing execution is thus the availability of data.

The operational semantics associated with data flow is usually represented as a simple directed graph. In a data flow program graph, each node represents a function. Arcs connecting nodes represent a uni-directional data path which carries a data token (e.g. a partial result) from a producer node to a consumer node. A node performs some operation, which is a function mapping inputs into outputs. In general, a node is enabled for execution when a data token is present on each of its input arcs. The node then executes and consumes the set of inputs, removing one data token from each arc. The inputs are processed according to the specified function and releases a set of result tokens on to the output arcs thus enabling further nodes. The data token plays a dual role, supporting both the data mechanism and the control mechanism; the flows of data and control are thus identical and inseparable.

In a simple data flow model of computation, data tokens are considered to move along the arcs of a data flow graph to the consumer nodes, the nodes being enabled for execution only when all of the input data are present. However, a restriction must be applied to the movement of tokens, such that only one token may be present on an arc at any given time. Without this restriction, it is not possible to identify which set of tokens form

the input to an operation at a node [WATS82]. A major disadvantage of this restriction is that iterative and recursive program structures cannot be supported (directly), which in turn unnecessarily limits the amount of parallelism that can be exploited in a program [SILV83, ARVI82]. This is sometimes referred to as the "static" data flow model because of the static nature of the data flow graph being implemented.

A number of schemes have been devised to overcome this deficiency [WATS82, ARVI82], which essentially require a label or tag to be carried with each data token identifying uniquely the context of that token. However, this adds significantly to the hardware complexity with a new requirement to perform a matching operation, of each tagged data token against other tagged data tokens to form a complete set. In the Manchester data flow machine, this operation is performed using a pseudo-associative store employing a hardware hashing technique [SILV83]. This general approach is usually termed the "dynamic tagged" data flow model, for obvious reasons. In a sense then, the dynamic model is just a generalization and extension of the static model.

The data flow concept supports very fine grained parallelism. If only one processor were in the system, all functions would execute on that one processor, with the result being the same as if it were run on a conventional computer with one processor. With more processors in the system, the functions are distributed amongst the available processors. Hence, the performance of the system may be tuned to requirements without re-writing the software. In data flow, the amount of hardware in the system is invisible to the programmer. Depending upon whether an application can use all available hardware, an n-fold improvement in throughput results from using n processors. This feature of data flow machines also results in hardware which is modular, and very tolerant of failures in an individual processing element, which are important system requirements (see section 2.2).

We believe that the concept of a data flow architecture could form the basis of a general-purpose optical computer, and also overcome a number of difficulties which have beset electronic data flow computers. The fine-grain parallelism found in data flow models is very well-suited to optical implementation, though it may be necessary to simplify the internal structure of the processing elements in comparison with the electronic equivalents. In the first instance, a program of development and research into an optical data flow computer should address the simpler "static" model in order to establish the operational principles. Later work could extend these ideas to a more general, dynamic tagged data flow model.

In both models there is a need for both global and dynamic interconnections, between processing elements and stored, enabled instruction "packets" (ie operation and operands). One method which might be suitable uses an array of independently controlled mirror elements as, for example, the CLAFLEX technique (also discussed in section 2.4.3) [WANG85]. In the dynamic model, there is an additional requirement to form sets of matching data tokens. Clearly, this problem will require investigation but a number of possible solutions occur to us, such as a matched-filter system based on that of Vander Lugt [VAND64] or perhaps a holographic technique. It should perhaps be noted that existing machines which implement a dynamic tagged data flow model [WATS82] have encountered problems in meeting expected performance which seem largely due to inefficient implementation of the matching function.

The potential applications of an optical data flow computer are very wide ranging (as discussed in section 2.3), and it seems fair to describe such a machine as a parallel general-purpose computer. In addition, the rapid progress being made in the development of Functional programming languages (see section 2.3) could lead to new levels of confidence in correct operation of very high performance systems.

2.4.3 Massively Parallel Architectures

A number of architectural proposals in recent years have made a fundamental assumption: that the number of processing nodes available to a given task is very large. Frequently, these proposals emerged as a result of the needs of those studying natural and artificial intelligence. These researchers are commonly faced with problems whose solution is most naturally suited to massively parallel models of computation. Sometimes, the term "connectionist" [FELD85] is mentioned in this context, to emphasize the point that many such models "store their long-term knowledge as the strengths of the connections between simple neuron-like processing elements" [ACKL85]. These models (sometimes called neural network models) were (originally) largely inspired by neurophysiology. An early example of this was the "perceptron" of Rosenblatt [ROSE62], in which the basic element is a model of a neuron, but which was found to have serious limitations [MINS68]. However, interest in the area was revived by Hopfield [HOPF82] who proposed a very simple model of a neuron having only two states. A network of such neurons is interconnected with various prescribed strengths or weights, such that the application of an external stimulus causes the neurons to adjust their states, in a manner which converges to the stored pattern most similar to the external stimulus. Essentially, this operation can be regarded as a nearest-neighbor search and is fundamental to the tasks of pattern recognition, associative memory, and error correction [FARH85]. A similar approach is found in "Boltzmann Machines" [ACKL85] which are described as a type of parallel constraint satisfaction network "capable of learning the underlying constraints that characterize a domain simply by being shown examples from the domain". A group at the University of Rochester [FELD85] is developing a parallel "connectionist" simulator, to be implemented on the Bolt, Beranek and Newman "Butterfly" multiprocessor [TREL86] using 128 processors.

A slightly different approach to the use of massive parallelism is suggested, in the implementation of semantic networks (see section 2.3). Fahlman [FAHL79] introduced a system called NETL, in which it is proposed that real-world knowledge is represented in the form of a semantic network realized directly in hardware, using a technique called "marker-propagation". More recently, Hillis at MIT has outlined a generic architecture which he calls a "Connection Machine" [HILL85]. This was originally developed to implement the marker-propagation algorithms required in the NETL system, but is in fact capable of much more flexible operation than this. Hillis suggests, for example, that the Hopfield model of a neural network fits well with the Connection Machine principles. The Connection Machine is perhaps the first hardware realization of massive parallelism, the prototype being constructed using 64K (i.e. 2^{16}) processors each having 4K (2^{12}) bits of memory and a simple bit-serial arithmetic-logic unit. The processors are connected using a packet-switched network in a binary n-cube (hypercube) topology, with an adaptive routing algorithm. The Connection Machine appears to have a modular construction and is potentially very tolerant of failure in single processors due to the huge amount of redundancy, both of which are important requirements in the system context (see section 2.2).

We believe that massively parallel architectures with the general properties described above are likely to benefit substantially from the advantages of an optical implementation. This view is supported in the work of Farhat et al [FARH85] with an optical implementation of the Hopfield model, and in the work of others such as Athale [ATHA85] and Fisher [FISH85]. Obviously, the highly parallel nature of these models and architectures is an important indication of the appropriateness of an optical implementation, but equally significant is the simplicity in the function of each processing element (or neural model) in these schemes.

Another way in which optical techniques support these architectures is the requirement for global interconnections between processing elements. This is often achieved using a fairly conventional optical vector-matrix

multiplier in the associative memory models. For an optical implementation of the type of architectures proposed by Hillis (i.e. Connection Machines), we seem to need global and dynamic interconnections between discrete processing elements, perhaps configured in a rectangular array. One possible approach to achieving such interconnections is the use of a corresponding array of independently controlled deformable mirror elements such as described in the CLAFLEX technique [WANG85]. A similar approach may lead to the implementation of a spatially-addressable parallel (multi-port) memory.

These massively parallel architectures could have many applications in current and future systems, both in the processing of sensor data at high speed, and in the concurrent processing of very large data structures. Systems employing pattern recognition techniques, for example vision and speech recognition systems, may usefully exploit the nearest-neighbor associative storage properties of the Hopfield approach. Hillis has suggested several quite general applications for the Connection Machine [HILL85], including image processing, VLSI simulation and semantic networks.

2.5 CONCLUDING REMARKS

We have shown that there are many important issues concerned with the design of complex systems. The general system requirements are seen to have implications in all aspects of a system, influencing the choice of architecture and hardware, as well as the high-level system software structure.

Parallel computing techniques have much to offer in terms of increased performance, but raise many questions concerning programming language styles, computational models and machine organisation. In the context of optics, other issues are still to be resolved, such as number representation, and optical "circuit" techniques.

Three candidate architectures were discussed and each of these seems to hold potential for development using optical or hybrid techniques. Systolic architectures seem most likely to yield positive results in the short term, although massively parallel and data flow architectures have the attraction of much broader application as a general-purpose parallel computer. Data flow architectures are capable of using their hardware resources most effectively and are inherently tolerant to hardware failures, but penalties are incurred in the volume of data communication required for a given computation. Massively parallel architectures such as the Connection Machine have perhaps the currently most important virtue for optical implementation, in their greatly simplified processing elements and emphasis on interconnections.

The most general conclusion we can draw, which is also a recommendation for further investigation, is that the machine (hardware) organization for an optical parallel computer should comprise processing elements of an ultra-simple nature and a mechanism to provide dynamic global interconnections between each element.

3. TECHNOLOGY

3.1 INTRODUCTION

3.1.1 Requirement for Optical Computing

Much attention has been focussed in the past on the research and development of analog techniques to perform processing functions. The use of the Fourier transform property of the lens, and the hologram in phase conjugation are just two areas which are being well researched.

The majority of current electronic computers are by contrast discrete in operation and nearly all computation functions are discretely based, with information often being transformed from an analog to a digital nature or vice versa at the inputs or outputs. There are many reasons why a digital computer has been so successfully developed. One of these is its ability to perform arithmetic functions adopting simple binary levels which offer immunity to accumulative noise. A consequence of this immunity is that no fundamental constraint is placed on the level of accuracy of computation. It could seem reasonable therefore to presume that an optical computer, with the objective of being able to process information at much higher rates than their electronic counterparts, would benefit from the noise immunity offered through the adoption of a discrete number representation. Accordingly, we have constrained our attention to technologies applicable to the development of an optical, digital computing capability.

To successfully adopt a discrete number representation the use of elements which display a non linear characteristic is essential. Additionally, any non linear characteristic should have the capability of being invoked through the application, or presence, of light energy. There exist many materials which display a non linear effect on light reflecting from or propagating through the material, but the majority of

them rely on an external influence such as an electric field or pressure, to cause the effect to occur. A material, or element which requires only the presence of light to effect a non linear characteristic on light is highly desirable. Considering available technologies which must ultimately source and detect the processed information, this characteristic should manifest itself as a phase, polarisation or preferably intensity modulation rather than say, a frequency modulation, which may lead to complications in design such as cascadability. The aim effectively is to use light not only as a carrier of information, but also as an instigator of a processing action.

A common factor of many of the architectures discussed in section 2 is the need for dynamic or global interconnect. Communication can often be a restrictive bottleneck in the development of computer systems with high processing rates. Light can provide the bandwidth required to realise high communication data rates but is limited often by the transport medium. Technologies suitable for realising dynamic or global interconnect are therefore considered and their operating requirements discussed.

3.1.2 Available Technologies for Discrete Optical Processing

Many materials have been identified which exhibit non linear optical characteristics. Further, some materials have exhibited optical nonlinearities of a nature which allows the phenomenon of optical bistability to be observed e.g. [PEY83], [DAGE84], [MILL81]. Optical bistability, where an input signal may result in one of two output levels, dependent on input signal history, is regarded as an important characteristic for optical computing.

A basic technology requirement, to allow the development of an optical digital computing capability, may be defined as follows:-

(a) The optical non linearity must be invoked through the application of light energy.

(b) The optical non linearity must occur at a wavelength readily provided by current laser sources.

(c) The magnitude of the non linearity must be sufficient to allow it to be usefully exploited at low optical power levels.

(d) The material, when incorporated in a device, allows the phenomenon of optical bistability to be observed.

GaAs, InSb, and ZnSe are a few materials that have shown sufficiently large refractive index non linearities, to be usefully exploited. When used as the spacer in a Fabry Perot interferometer, the devices exhibit optical bistability at visible or infra red wavelengths and at low power levels. Such devices have been fabricated, by, most notably perhaps Heriot Watt University, Edinburgh and the University of Arizona, Tucson.

We have therefore concentrated on the use of non-linear Fabry Perot interferometers as basic building blocks in the development of an all optical computer.

3.1.3 Principal Technology Objectives

The provision of the basic building block does not imply that the development of an optical computer is imminent. Many advances and developments, both technological and architectural, must be made before worthwhile computing functions and rates are realised. To accelerate progress towards the achievement of such a goal we adopted the following objectives:

(a) To become familiar and conversent with optical technologies in general and optical bistable interferometers specifically.

(b) To identify present performance limitations and strengths leading to the development of engineering specifications.

(c) To identify and investigate techniques, based on the use of light as signal and control inputs, for the development of basic logic functions.

(d) To identify and investigate the role of bistable interferometer technology in optical digital computing.

In section 3 we discuss the non-linear interferometer approach to optical computing. We define the functional operating requirements for non-linear interferometers, based on the published performance of those developed by Heriot Watt University, Edinburgh. We then proceed to identify the operating requirements of the supporting technologies necessary to fully exploit the interferometer characteristics. We conclude with a discussion of the current performance and identify areas which require further investigation and optimisation before a practical system may be realised.

3.2 INTERFEROMETER APPROACH TO OPTICAL COMPUTING

3.2.1 Description of Characteristic/Operation

A more detailed discussion of the theory of operation of Fabry Perot bistable interferometers where the bistability is thermally induced is given elsewhere, e.g. [JANO85]. A brief description of its operation is given below.

The characteristics of optically bistable interferometers are determined primarily by the non linear characteristic of the cavity material. ZnSe exhibits a non linear thermally induced refractive index variation, due to the absorption of incident light. When used as the cavity material in a Fabry Perot interferometer the refractive index change, equating to an optical path length change, alters the reasonant wavelength of the cavity.

If the interferometer is designed such that the optical path length does not equal an integral number of wavelengths at the wavelength of incident light , it is said to be off resonance. As the incident intensity is increased, more energy is absorbed leading to greater heating of the material, and hence longer optical path lengths due to the positive increase in refractive index. Hence, with increasing intensity, the interferometer path length approachs that required for resonance at the input wavelength. Positive feedback is introduced via the mirrors causing the cavity intensity to build up.

Once the resonant condition has been satisfied, any reduction in input intensity may not immediately result in the interferometer falling back to its off resonant condition as much less power is required to maintain the resonant condition than that required to achieve it in the first instance. Clearly the role of the substrate which acts as a heat sink is particularly important in establishing this characteristic.

Figure 3.1 [JANO85] shows the range of transmission characteristics that are obtainable by varying the initial detuning of the interferometer. The initial detuning is achieved through a change in the optical thickness of the interferometer by altering the angle between the incident beam and the plane of the interferometer.

It is as a result of these characteristics and their capability for performing logic functions that we have focussed our attention on determining the role of the non linear interferometer in optical computing.

3.2.2 Example of Simple Logic Functions

The characteristics exhibited by the interferometer rely on the absorption of energy derived from the input beam. Due to the material absorption characteristic being coherence insensitive, all of the characteristics shown in Figure 3.1 can be obtained using 2 or more input beams. Only 1 beam will resonate however and it is this beam that will define the useful transmission and reflection powers of the interferometer. This allows the necessary control to be introduced and thus logic functions to be developed.

3.2.2.1 AND/NAND Gate

A multiple input AND/NAND gate function may be achieved simply through the appropriate biasing of the holding beam. The holding beam is adjusted to a level where only the application of all inputs 'high' will provide sufficient energy to cause the interferometer to switch on resonance. The AND function is obtained from the transmitted beam and the NAND function from the reflected beam.

3.2.2.2 OR/NOR Gate

The OR/NOR function is similarly achieved.

Through appropriate biasing of the holding beam the application of any one signal beam may be sufficient to create resonance. The OR function is obtained from the transmitted beam and the NOR function from the reflected beam.

3.2.2.3 Fan Out

Adoption of the above approach to biasing to alter functionality will result in a variety of output power levels. The ability of devices to cascade is of fundamental importance in the development of any realistic worthwhile processing function. A variation in holding powers will result in a variation in output powers and hence devices will have a variety of fan out capabilities. It may therefore be prudent to insert attenuators into the signal beam paths such that the AND and OR gates have identical holding beam powers thereby simplifying both generation of holding beams and fan out considerations. Further considerations to this topic is given in section 3.3.

3.2.2.4 Memory

Due to the existence of positive feedback characteristics wide hysteresis loops are obtainable. This characteristic allows a memory function to be developed by biasing the holding beam within the hysteresis loop such that the application of a signal beam will cause the resonant condition to be achieved. On removal of the signal beam, the feedback mechanism allows the resonant condition to be maintained and hence the interferometer 'remembers' the previous input condition.

3.2.2.5 More Complex Functions

More complex functions which may be considered useful or necessary for optical computing are the exclusive - OR/NOR function and the full adder.

The exclusive OR function, demonstrated by Smith et al [SMIT85] required the use of two interferometers. The experimental layout used by Smith is shown in Figure 3.2(a) and the output characteristics are shown in Figure 3.2(b). Note that both interferometers require to be operated in their reflection mode and that for highly parallel applications two spatially distinct interferometers may be preferred to implement this function.

The full adder function, demonstrated by Wherrett [WHER85] illustrates the functional capability of a single interferometer by using both transmitted and reflected beams for the carry and sum respectively. (Figure 3.3 refers). Wherrett demonstrates that a reduction, for a full adder, of 14 to 1 where the electronic equivalent is realised with NOR gates only, is possible.

It is interesting to note that in developing such a function, an exclusive-or and exclusive-nor function was exploited in the reflection and transmission characteristics. Hence a single interferometer exclusive -or/nor function is realisable. Figure 3.2 illustrates the characteristic and function. This characteristic is further discussed in section 3.5. The demonstration of the remaining ten commutative logic functions have not been considered in this report.

3.2.3 Consideration of Architectures

With the necessary building blocks in terms of logical functions being demonstrated, consideration must be given to what topology/architecture would best exploit their strengths.

The main strength of non linear interferometers based on ZnSe lies not with their switching speed, which by current electronic standards is very slow but in their capacity for supporting immense parallelism. (Advances in switching speeds are envisaged, through optimisation of materials and device construction).

It has been suggested by Smith et al [SMIT85] that an array of 10^4 individual interferometers are capable of being operated simultaneously on a single substrate measuring only a single square centimetre. Clearly the potential for providing exceptionally high processing rates lie with marrying a slow but acceptable switching speed of a few tens of

microseconds to the parallelism capability to perform gigabit operations per second (GOPs). Highly parallel architectures therefore may offer the most suitable structure for exploiting the interferometer technology capabilities.

Having determined the degree of parallelism possible, it is clear that supporting technologies will need to be identified in its realization. It is not envisaged that a single plane of interferometers will be able to provide the processing power needed, to gain any substantial advantage over electronics, but the combination or sequence of planes that will do so. Therefore large numbers of beams must be generated to supply holding power, routed to the relevant positions/interferometers, and ultimately detected.

It is necessary to first define the functional operational characteristics and requirements of the interferometer technology and architecture before the identification and subsequent specification of the required supporting technologies.

3.3 INTERFEROMETER INVESTIGATION

3.3.1 Requirement for an Engineering Specification

To successfully operate a single array of interferometers requires the assistance of many supporting technologies. An array of holding beams will need to be generated, to distribute power to each individual element of the interferometer array. A laser source is required sufficiently powerful to satisfy the interferometer power requirements and a detector array is required to detect the processed signals. In configurations where a fan out of 2 or more is required deflecting and focusing elements will need to be defined.

There are many considerations in the definition of supporting technologies; wavelength and polarisation, dynamic or fixed interconnect, pulsed or cw operation etc, which must all satisfy the operating requirements of the most important building block that provides the processing function. It is important therefore to comprehensively define the functional operating requirements of the interferometers before proceeding to specify the required characteristics of the supporting technologies.

An additional benefit of determining such a specification is perceived in the provision of a concise description of the state of the art performance levels of the technology, allowing identification of areas which could most benefit from optimisation, and the associated trade offs involved. Further, this approach allows the early definition and evolvement of the operational and interface standards discussed in section 2.2.3.

Thus the interferometer functional operating (engineering) specification is defined with consideration given to its impact on design.

3.3.2 Definition of Engineering Specification

The characteristics and dependencies discussed in this section are based upon the performance of Fabry Perot interferometers developed by Heriot Watt University, Edinburgh. The interferometers use Zinc Selenide as the active cavity material and Zinc Sulphide and Thorium Fluoride in the mirror construction. The major factors determining performance, from an engineering consideration, are discussed, for example, switching power variation with spot diameter and the variation of switching time with spot diameter. We then go on to define an operational engineering specification for the bistable and transphasor characteristics displayed by the interferometers.

3.3.2.1 Characteristic Dependencies

The following characteristics and their dependencies are discussed:

The dependence of switching power on spot diameter.

The dependence of switching time on spot diameter.

The effects of an interferometer on a Gaussian beam.

Power transfer characteristics.

The dependence of the substrate thermal conductivity on switching power and switching time.

3.3.2.2 The Dependence of Switching Power on Spot Diameter

Figure 3.4 illustrates how switching power, defined as the minimum input power necessary to cause resonance, decreases linearly with spot diameter. It can be seen that for maximum power efficiency, the minimum spot diameter possible should be used, and thus diffraction limited spot sizes are optimum.

3.3.2.3 The Dependence of Switching Time on Spot Diameter

The dependence of switching time on spot diameter is shown in Figure 3.5. Further investigation is required to more accurately determine the characteristic but the trend does indicate that the minimum switching time is achieved with minimum spot diameters.

Also included on Figure 3.5 is an indication of the effect of switching the device with an input power greater than its switching power (overdriving). It can be seen that by overdriving, the switching time can be reduced.

This conflicting requirement is discussed further in section 3.3.5.

3.3.2.4 The Effects of an Interferometer on a Gaussian Beam

Beam divergence, shown in Figure 3.6, has been observed in bulk ZnSe due to self focusing [TAGH85]. Further, it has been reported that when on resonance, defocusing has been observed leading to a ring structure of the far field transmitted beam, and suggested that the useful power level of the switched on transmission can actually be less than that of the switched off transmission [WHER84]. The effect of beam divergence on cascadability of interferometers, in transmission, is therefore an important area requiring further investigation.

3.3.2.5 Power Transfer Characteristics

Figures 3.7 and 3.8 illustrate the difference in power transfer characteristics for reflection and transmission for an interferometer tuned for bistability. As can be seen the power transfer from incident to transmitted power is very low at approximately 5%. This places an important consideration in the development of engineering specifications for cascadability.

3.3.2.6 The Dependence of Substrate Thermal Conductivity on Switching Power and Switching Time

The mechanism by which the optical nonlinearity occurs in ZnSe is thermo-absorptive. It is the absorption of photons, and the subsequent generation of phonons which heats the lattice and alters the refractive index. If the thermal conductivity of the substrate is too great, then the necessary temperature rise will only occur, at high incident powers. If on the other hand, the thermal conductivity is too low, then switching will occur at very low indicent power levels but the heat generated, unable to be dissipated, may distort the structure. Further investigation is required to identify the trends and trade offs associated with this dependence.

There exists other dependencies, (eg variation of power tranfer characteristics with cavity length etc) which are believed to be of less importance, in terms of engineering specifications definition, and therefore not discussed here.

3.3.2.7 <u>The Device Family Characteristics</u>

Figure 3.1 illustrates the family of transmission characteristics obtainable from a non linear interferometer, by varying the initial detuning from resonance. The three characteristics are referred to as:-

1. Bistable

2. Transphasor

3. Limiter

Detuning the interferometer from resonance is achieved through altering the angle of incidence of the interferometer to the incident light beam.

The characteristics shown are typical of an interferometer illuminated such that the incident beam propagates through the interferometer before the substrate.

The operational requirements of the transphasor and bistable characteristics, based on an OR function, are discussed in sections 3.3.2.8 and 3.3.2.9. The limiter characteristic has not been identified of performing a useful processing function and is therefore not discussed in detail in this report.

3.3.2.8 Bistable/Transphasor Operational Requirements

The application of an interferometer to perform a binary logic function, operating with continuous wave power source and resulting signals, is considered. In any practical application of these devices within a circuit, various factors will determine tolerance and variations on the parameters which dictate performance. A fundamental requirement is for each element to perform reliably and unambiguously. Consequently, these elements must be designed with sufficient margins in order to accommodate these imperfections.

The following analysis relates to a circuit configuration in which an interferometer is driven by two signal beams (a minimum requirement to perform a gate function), and a holding beam of fixed nominal intensity held at a level below the switching level required to obtain resonance. The signal sources are assumed to be two similar bistable devices where both are operating in transmission mode, or both are in reflection mode. Additionally, each source is required to drive two similar interferometer devices (i.e. a fan out of 2, identified as a minimum requirement to obtain outputs from a looped configuration). The interferometer being driven can either be a transphasor performing an asynchronous gate function, or a bistable device operating as a gated synchronous latch memory.

With reference to Figures 3.7 and 3.8, let:

P1 = PT1 or PR1, at nominal value

P0 = PT0 or PR0, at nominal value

Pu = reference 'power up' switching power, assumed to be constant for the purpose of analysis (similarly defined for transphasor characteristic)

Ph = holding power, at nominal value

For the driven device to operate as an AND/NAND logic function, the critical switching equation to obtain resonance is:

$$P1 + Ph > Pu \qquad (1)$$

and to avoid resonance:

$$(P1 + P0)/2 + Ph < Pu \qquad (2)$$

For unambiguous logical operation:

$$P1 > P0$$

Also $P1$ or $P0 < Ph$

Hence, $Pu - P1/2 > Ph > Pu - P1 \qquad (3)$

Within the boundary conditions of (3), and inspection of Figure 3.7, PT1 max and PTO max may be substituted within the equation (1) and (2) to derive a tolerance on the largest variable Ph as 57.49 mW $\pm$ 0.6%.

By similar deduction for the device operating as an OR/NOR logic function, the critical switching equation resonance is:

$$(P1 + P0)/2 + Ph > Pu \qquad (4)$$

and to avoid resonance:

$$P0 + Ph < Pu \qquad (5)$$

In this case, the rquirement for Ph is 58.18mW $\pm$ 0.59%.

This indicates that the OR/NOR function is more critical, so this requirement is used for further analysis.

Interaction between adjacent spot positions within an etalon can create a crosstalk effect which results in an apparent increase (by addition) to the level of Ph. For a spot separation of four spot diameters this can change the switching threshold by 15% [TAI82]. In addition, the following variables have been suggested [TOOL86]:

Laser diode $\pm$ 1%

Beam Array Uniformity $\pm$ 2%

Interferometer switching threshold variation $\pm$ 2%

These values indicate the following tolerances within the critical switching equations:

P1 or P0 $\pm$ 3%

Ph +(5+m)%, -5%
where m corresponds to 'crosstalk'

Substituting worst case tolerances into (4) and (5) produces:

$$0.485\ (P1 + P0) + 0.95\ Ph > Pu \qquad (6)$$

$$1.03P0 + (1.05 + m/100)Ph < Pu \qquad (7)$$

By considering boundary conditions within (6) and (7) the following deductions can be made:

$$\text{Power Transfer Efficiency } P1/PH > \frac{(0.1 + m/100)}{(0.485 - 0.545/n)} \tag{8}$$

$$\text{Contrast Ratio } n > \frac{0.545}{0.485 - (0.1 + m/100)} \tag{9}$$

$$n = P1/PO$$

Equations (8) and (9) above describe the lowest limits for Power Transfer Ratio and Contrast Ratio, and both conditions must be satisfied simultaneously by an interferometer design to operate as an OR/NOR function. Equations (8) and (9) can be used to assess interferometer suitability by observation of the transfer characteristic curves. Used in conjunction with (6) and (7) the optimum position for Ph nominal can be located. It should be noted that no account has been taken for possible additional causes of variations, such as mirror reflection losses and attenuation in the transmission path between the logical devices. Consequently, the equations must be considered as optimistic, and further contingencies should be considered.

The above equations apply equally well to a transphasor or a bistable device, as the derivation took no account of the shape of the transfer characteristic curves.

For a bistable device a further condition must be met to ensure that the memory state is retained:

$$0.97\,PO + 0.95\,Ph > Pd \tag{10}$$

Where Pd is the switching power at "power down". Providing equations (6) to (10) are satisfied with suitable contingencies then in order to conserve power, Ph should be positioned as low as possible. However the switching speed to achieve on resonance is dependent upon the total incident power at the time of switching. Consequently the level of Ph may need to be positioned higher than the minimum levels determined by the above equations.

3.3.2.9 Review of Performance and Specification

Table 3.3.2 describes an interferometer device engineering specification format, which should be used to explicity define limits of performance for the stated function. The values within this table are for illustration only, as information is not available to verify all figures quoted. It should be noted that critical tolerance limits are defined to provide the user with all necessary information for designing the circuit.

Section 3.3.2.8 describes a minimum requirement for the operation of a two input OR/NOR logical function. Equations (6) to (9) detail necessary limits for the signal and holding beams in relation to the switching power for obtaining resonance. In addition equation (10) describes a further limit if the device performs as a bistable element.

If crosstalk equals 15%, then from (9) the Contrast Ratio P1/P0 must be greater than 2.32. By observation of the bistable characteristics, this condition can be satisfied for transmission (Figure 3.7), and is never satisfied for reflection (Figure 3.8) for a particular value of Ph.

For the transmission characteristics, the maximum value of Contrast Ratio is approximately 4.72. Substituting this value into (8) produces a Power Transfer Ratio of 0.676 (indicating that PT1 must exceed 24 mW for Ph at 35mW).

Consequently both the transmission and reflection characteristics are inadequate to meet the requirements for successful operation.

Suppose the first approach to optimisation is to reduce the crosstalk to 1% (possibly by separating the spot centres further, or by physical pixellation of the etalons to provide barriers between spots). Using (9) the contrast Ratio must exceed 1.45. Substituting a Contrast Ratio of 4.72 into (8) for the transmission characteristic produces a minimum requirement of 0.3 for Power Transfer Ratio (i.e. PT1 greater than 10.5mW at Ph of 35mW).

A Contrast Ratio of 1.45 appears achievable on the reflection characteristic for Ph at 35 mW. However for these values, PR1 must be greater than Ph and this is not achievable.

A similar assessment of the transphasor characteristic curves shown in Figure 3.9, reveals similar inadequacies in performance.

From the above discussion, it is clear that the configuration and requirements stated in section 3.3.2.8 cannot be met by the interferometers characterised in this document. It can be seen from (8) that an increase in Contrast Ratio permits a reduction in Power Transfer Ratio. It is suggested that improvements for the characteristics discussed may be realised if the reflection efficiency is reduced and the transmission efficiency is increased, together with the introduction into the fabrication process of suitable techniques for spot to spot isolation. Further to these steps consideration can be given to optimising interferometer design to perform in transmission only or reflection only.

3.3.3 Optimisation Considerations

The engineering specification, defined for the bistable characteristic of the interferometer, table 3.3.2 provides a concise source of relevant operating characteristics and conditions. The figures contained within the table are generally typical and have not been optimised for any particular characteristic. It is intended to regularly update and expand this engineering specification, to reflect achievements in technology design and optimisation of functional characteristics, to provide a clear statement of the state-of-the-art and thus provide the necessary information for assessment of practicality of design.

It is anticipated that due to a number of optimisations with contradictory trade offs (switching power vs. switching time, for example see section 3.3.2.1) a number of devices will evolve displaying individually optimised characteristics.

Table 3.3.3 identifies target performance requirements.

Many areas have been identified that would benefit from optimisation and could have an effect on the practicalities of any proposed design and implementation [WHER84, WHER86]. The primary considerations for optimising device performance in engineering terms are now discussed.

3.3.3.1 Power Transfer Characteristics

Ideally, the power level at which switching occurs of any optical logic function ought to be high enough to provide sufficient noise immunity but sufficiently low to ease the total power requirement on the laser source. Unfortunately the present non linear interferometer characteristics exhibit switching power levels in the 30 to 40mW range; well above the level of any noise source currently envisaged to exist in an 'optical computer' environment. Thus one area which would provide immediate rewards in terms of supporting technology requirements through optimisation, is a reduction in switching power levels.

Non linear interferometers provide two functional characteristics simultaneously in normal operation: that defined by the reflected beam and that defined by the transmitted beam. There may exist many situations where both characteristics are not required simultaneously from the same array of interferometers. It may be acceptable therefore to optimise for either the reflected characteristic (NAND/NOR) or the transmitted characteristic (AND/OR) but not for both simultaneously.

3.3.3.2 Reflection Characteristic

If the characteristic desired is obtained from the reflection rather than the transmitted beam then any power transmitted from the back face of the interferometer can be regarded as a loss and thus should be minimised. Increasing the reflectivity of the back face to unity will result in a reduction of input intensity necessary to cause resonance. [WHER84].

Additionally, setting the back face reflectivity to unity also optimises the difference in reflected power between on and off resonance. This is clearly highly desirable from fan out considerations.

3.3.3.3 Transmission Characteristics

Similarly if the characteristics desired, (AND/OR), are obtained from the transmitted beam, then any power reflected from the cavity can be regarded as a loss and thus must be minimised. This suggests the reflectivity of the front face should approach unity, but by doing so less power is transmitted into the cavity in the first instance and hence will result in an <u>increase</u> in incident intensity necessary to cause resonance, a conflicting requirement.

A possible solution to be investigated, is to develop a mirror structure which displays different reflection coefficients dependent on the direction of propagating light. This may be achieved through refractive index profiling or the use of antireflection coatings.

A further improvement may be obtained through altering the structure of the Fabry Perot interferometer itself. Current structures consist of $ZnS/ThFl_4$ mirror stacks with the active medium acting as the cavity. As the non-linearity is invoked through the absorption of light the efficiency, defined as the ratio of transmitted power to input power, may be increased by reducing the length of absorbing medium. To maintain the finesse of the cavity necessary for bistable operation, the optimum cavity length must be maintained [WHER84]. A structure in which the ZnSe active medium is contained in the mirror stacks may satisfy the finesse condition while reducing the length of absorbing material. Investigations into this structure are currently being undertaken by Heriot Watt University, Edinburgh.

3.3.3.4 Switching Characteristics

By electronic standards, the switching speed of non linear interferometers, using a thermal-absorptive change in refractive index to determine resonance is very slow. An analysis of the switching mechanism [GOLD81] has shown the total switching time from off resonance to on resonance is determined by two components; the time taken for the optical cavity length to reach the necessary integral number of wavelengths and the time taken for the cavity intensity due to resonance to build up. (The switching time from on resonance to off resonance is determined by the thermal conductivity of the substrate). Goldstone has shown [GOLD81] that through the application of a signal power in excess of the minimum power required to induce switching, the time taken to achieve the correct optical cavity length can be reduced such that the intensity build up time dominates. While providing the mechanism to reduce

switching times, it is at the expense of higher input power. However, input power needs to be considerably reduced in practice. It may transpire that the switching time of the non-linear interferometers can be tailored for particular applications with due consideration to power sourcing capabilities in future.

A benefit of this phenomenon, known as critical slowing down, is the increase in noise immunity offered. Due to the time taken to switch, any transient effects on signal beam may not be of sufficient duration to cause switching. This will allow holding beam powers to be biased much closer to the operating point of the characteristic than present noise sources may suggest, whilst ensuring stability in the off resonance condition.

The switching time from the on resonance state to the off resonance state is determined by the thermal conductivity of the substrate. This switching time limits the maximum switching frequency obtainable from present interferometers. The substrate also plays an important role in determining the power transfer characteristic and in providing a heat sink capability, necessary in the dissipation of heat energy generated by an array of 10^4 elements simultaneously operated. It is therefore identified as a area worthy of further investigation.

3.3.3.5 Incident Angle of Operation

The incident angle, determining the interferometer power transfer characteristic, by altering the level of initial detuning, is an important parameter in the engineering specification. It is envisaged that in most practical systems, beam routing will be necessary simply to provide feedback or to avoid "long" sequences of arrays. This beam routing, traditionally performed by using extremely flat and efficient mirrors, may be achieved through altering the angle of incidence between the interferometer and the input beams.

By careful consideration of the optical geometries, it is likely that the number of mirrors for a practical system may be reduced. However, this may require an increase in the range of incident angles for which useful operation of the non linear interferometers is defined.

It is at present unclear as to the range of incident angles that may be achieved by careful design of the interferometer for any given characteristic. Further work is required to investigate this property and the flexibility it may provide. One can envisage a range of devices each optimised for a particular angle of incidence and for a particular characteristic.

3.4 SUPPORTING TECHNOLOGIES

3.4.1 Laser Sources

Laboratory prototype demonstrator sysstems, based on ZnSe interferometer technology, rely on an Argon ion laser system to provide the necessary coherent optical power. As is well known, powerful Argon ion laser systems are highly inefficient, consuming kilowatts of electrical power, and also very large. This makes them unsuitable for inclusion in any practical optical computer system. Alternative sources are therefore required which are compact, efficient and convenient to use.

Laser source requirements have therefore been defined for current and potential system needs, based on the interferometer operating requirements defined in Table 3.4.1.

Technology has yet to advance to a position where solid state laser diodes, lasing at 514nm, have been demonstrated. In contrast, laser diodes with an output wavelength of 850nm are readily available. Available output powers are not however sufficient at present to satisfy the CW total holding beam power requirements of current ZnSe interferometers, from one laser diode. (Note: Optical bistability, using a pulsed laser diode source, was first reported by Tarng et at [TARN84]).

Emphasis is placed therefore on the reduction of switching energies and the increase in power transfer efficiencies of ZnSe interferometers and in demonstrating CW optical bistability at near infra red wavelengths.

3.4.2 Detectors

Optical computers generate optical outputs i.e. the output signal is in the form of coded light intensities. These signals may require to be converted to electronic signals if further processing or electronic storage is required, for example, if the optical computer system is embedded in an electronic system.

Fortunately a large range of components e.g. photodiodes, phototransistors etc., are available at present which have sufficient sensitivity and responsivity to detect the targeted optical powers.

More problematic however is the detection of an array of optical signals. Using a lens system, the pixel density can be reduced at the output stage thus relaxing the requirement. Solid state image sensor arrays are identified as providing the necessary pixel density and resolution to satisfy present performance e.g. Reticon RA100X100, but further development is required to satisfy the target specification of 86×10^3 pixels/cm^2.

3.4.3 Routing and Generation of Beam Array

3.4.3.1 Introduction

With microelectronic circuits, as the performance and complexity increases, so does the number of interconnections. Ultimately, the interconnection problem dominates over other limiting factors.

To exploit the benefits of non linear interferometer technology, the requirements for cascading two interferometers and the provision of an array of holding beams represent interconnection problems that must be solved.

Solution of the light beam interconnection problem should be less limiting than corresponding multi-layer conductive interconnection path problems, encountered in conventional integrated circuit layouts. The reduction of crosstalk and the ability to intersect light beams without interference helps to reduce optical interconnection network restrictions.

3.4.3.2 Beam Array and Interconnection Requirement

A fundamental requirement for the operation of non linear interferometers is the formation of a beam array to provide holding beam power to each pixel. Similarly, to transport an array of beams requires the provision of focusing and routing elements. The ability to dynamically vary the routing of beams between pixels may also be an important requirement. Possible techniques to provide static and dynamic interconnections are discussed below.

3.4.3.3 Static Interconnections

Holographic techniques and components are attractive for this application due to their power efficiency and compactness.

A schematic diagram is shown in Figure 3.10 [JENK83] of the optical system for direct implementation of space variant connections. The gate outputs are imaged on to the interconnection hologram. This hologram consists of an array of subholograms, one sub hologram for each gate output, i.e. the holographic plate, is "pixellated" to correspond to the gate input. This hologram is encoded in the the Fourier transform plane, (i.e. wavefront) and a Fourier transform is taken optically (lens 2) to obtain the gate inputs in the image plane (Note only one of the pair of images is used, the other can be a test probe). The interconnections are formed in one particular diffraction order (e.g. 0), but in general, multiple diffraction orders are produced, the higher order diffractions being regarded as a loss. Lens 1 and Lens 2 may also be replaced by suitable holograms [CLOS75].

The hologram can be computer generated to implement this interconnection pattern [LEE70, JEN83].

Requirement for Holographic Optical Elements (HOE's)

The ability to computer generate Fourier plane holograms [LEE70] coupled with the low dispersion and low aberrations of multiple hologram optical elements [LATT72] makes them a very attractive combination for this application.

Other advantages are that

(a) HOE's can be fabricated into stacks of thin films, allowing optical elements to overlap if required, and

(b) are potentially inexpensive and simple to produce.

In defining the requirement specification for holographic optical elements, Table 3.4.3.3, dichromated gelatin holograms are specifically considered due to that versatility [CHAN80].

3.4.3.4 Dynamic Interconnections

The ability to alter the interconnections between interferometers is possible using Spatial Light Modulators. Spatial Light Modulators offer an alternative interconnection technology to holograms which could prove cost effective for laboratory prototype systems, where it is desirable to experiment with different interconnection schemes.

Additional uses for Spatial Light Modulators include the following functions:-

(a) Incoherent to coherent conversion.

(b) Wavelength to wavelength conversion.

(c) Serial to parallel/parallel to serial conversion.

(d) Real time holography.

(e) Long term storage of holograms.

3.4.3.5 Requirement

A general requirement specification for a Spatial Light Modulator was compiled, reference Table 3.4.3.5.

The frequencies of operation for the present and target specification requirement were chosen so as not to limit the input/output rate determined by the interferometer switching speed.

Write and erase time requirements were estimated to be consistent with operating frequency.

The storage time requirement was a compromise between an infinite storage time and a single clock cycle time. However the requirement of 1 hour is intended only a a guideline.

There exist many types of spatial light modulators and many more are being researched and developed. A review of spatial light modulators is given elsewhere (see [FISH86]), and this is clearly an area which requires further investigation.

3.5 STRUCTURAL REQUIREMENTS FOR IMPLEMENTATION

3.5.1 Etalon Approach

An advantage of using an interferometer technology, whose characteristics can be tailored for a particular angle of incidence, is that an optimised geometric configuration may be developed, for the function cell.

Using a miniature optical bench the prototype function cell can be constructed and its performance evaluated. It is then a matter of production effort to produce a prototype system with a full complement of logic cells.

The following points must be considered in detail in the design stage.

(a) Requirement for solid state laser sources to meet the holding beam power requirements.

(b) Requirement for holographic optical interconnect plates.

(c) Requirement for a suitable output photodetector matrix.

(d) Requirement for a suitable spatial light modulator and input devices.

The two obstacles to be overcome before production of the operational system are:-

1. The specification of the solid state laser source. An argon ion laser is only useful for bench evaluation of a prototype system. The final production version will use solid state laser sources.

2. The specification of the optical interconnect elements. For logic cells requiring holographic optical elements to achieve a high interconnection density, the design and development of the holographic interconnection plates must be proven for the production geometric configuration.

The ZnSe interferometer approach to an optical processor does not seem to adapt to monolithic integration as well as other technologies but there may exist the opportunity to adopt a hybrid approach by combining alternative technologies on the substrate.

In particular the following two combinations seem attractive:-

1. Integrating an output stage interferometer with a self scanning image sensing photodetector array.

2. Integrating the holding beam power source with each interferometer; by fabricating an array of solid state laser diodes and microlenses for each pixel on to the face of the interferometer.

It is important to ensure that all optical elements are mounted securely, and that external vibrations do not change the performance of the optical system for example, if the interferometer changes position relative to the source, then the initial detuning may vary and therefore the interferometer may not respond to the incident light signals in the desired manner.

Similarly if for example a micro lens vibrated, then the holding beam may overlap pixels and induce a transient error in the system. Clearly the optical system must be set up with precision and this maintained throughout system life.

A technique sometimes employed is called "minibench" and involves pouring a U.V. curable transparent epoxy resin between the optical elements in a system, and when the system is correctly aligned to cure the resin with U.V. light, thus 'setting' the positions. This system, whilst providing an effective "seal" and moisture barrier for the optical system, prevents maintenance by component replacement.

It is very important that the elements in the system have a uniform performance. Applying this criteria for example, to sources, requires the matched laser diodes, or feedback control electronics to provide uniform intensity to avoid any thresholding errors in subsequent interferometer stages. This consideration applies to the holograms, photodetectors, interferometers, indeed all optical components.

BISTABLE ENGINEERING SPECIFICATION

PARAMETER		TRANSMISSION			REFLECTION			
		MIN	TYP	MAX	MIN	TYP	MAX	
POUT '1'	"HIGH" LEVEL POWER OUTPUT	2.25	2.27	-	45.2	45.7	-	mW
POUT '0'	"LOW" LEVEL POWER OUTPUT	-	0.79	0.80	-	35.7	36.1	mW
PIN '1'	"HIGH" LEVEL POWER INPUT	2.22	2.27	-	44.8	45.7	-	mW
PIN '0'	"LOW" LEVEL POWER INPUT	-	0.79	0.81	-	35.7	36.4	mW
P U	"UP" SWITCHING POWER INPUT	59.5	-	59.5	59.5	-	59.5	mW
P D	"DOWN" SWITCHING POWER INPUT	-	34.0	35.7	-	34.0	37.5	mW
P WIDTH	BISTABLE LOOP WIDTH POWER INPUT	23.8	25.5	-	23.8	25.5	-	mW
SPOT SIZE (DIAMETER)		-	35	-	-	35	-	µm
SPOT-SPOT SEPARATION		700	-	-	700	-	-	µm
t_U	"UP" TRANSITION TIME	-	50	-	75	-		µsec
t_D	"DOWN" TRANSITION TIME	-	75	-	50			µsec
f_{MAX}	MAX. SWITCHING FREQUENCY	-	2.5	-	-	2.5	-	KHz
L^o	ANGLE OF INCIDENCE OF ETALON	2	-	+2	-2	-	+2	deg.
λ	OPERATING WAVELENGTH	-	514.5	-	-	514.5	-	nm
OPERATING TEMPERATURE		-	27	70	-	27	70	degC
P HOLD	HOLDING BEAM POWER INPUT	55.2	55.8	56.9	55.2	55.8	56.9	mW

Operating Conditions:
Max. fan-out 1
Max. tolerances: Ph ± 1%
PIN ± 2%

TABLE 3.3.2

BISTABLE INTERFEROMETER OPERATING CONDITIONS

	PRESENT PERFORMANCE	TARGET SPECIFICATION	WISHTIMATES
Operating Wavelength λ	514nm	850nm (1)	850nm (1)
Spot Size (Diameter)	35µm (2)	8.5µm (2)	8.5µm (2)
Spot Separation (Centre to Centre)	140µm	35µm	20µm
Calculated Pixel Density (No/CM^2)	5.1×10^3	86×10^3	250×10^3
Switching Power	Pwr Up. 60mW; Pwr. Dn 35mW (3)	Pwr Up. 20mW; Pwr. Dn. 8mW	Pwr Up. 0.1mW; Pwr.Dn 0.04mW
Total Incident Switching Power (per cm^2)	306W	1720W	25W (5)
Reflection Efficiency	'0' Min. 86%; 1 Typ. 63%; 1Max. 66%	60%	60%
Transmission Efficiency	'0' Min 1.1%; 1 Typ 4.4%; 1Max 3.7% (4)	60%	60%
Switching Time	-	Tav 0.2m sec	Tav. 0.02m sec
Switching Frequency		50KHz	50KHz
Etalon Size	1cm^2	1cm^2	4cm^2

1 Wavelength chosen to optimise the use of solid state devices

2 Criteria for choosing the spot sizes.
Present capability based on reported results.
Target and Upper Spec Limit : Minimum practicable spot diameter (diffraction limited)

3 Data reference Figure 3.3.4

4 Data reference Figure 3.3.4. Note present transmission efficiencies insufficient to allow operation with fan out and fan in of two

5 Based on maximum conducted heat dissipation : 7.5W/cm^2

TABLE 3.3.3

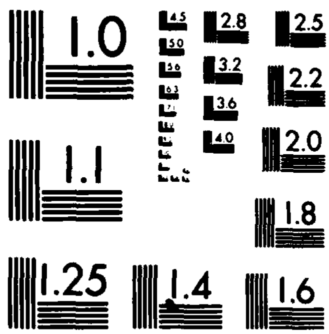
LASER DIODE REQUIREMENT

	PRESENT SPECIFICATION	TARGET SPECIFICATION
Total Power Output per Interferometer	306W	1720W
Pixellated area of illumination	$1 cm^2$	$1 cm^2$
Switching Frequency (minimum)	1KHz	143KHz
No. of Diodes allowed per etalon	9 (3 x 3 array)	16 (4 x 4 array)
Power Output per laser diode array	34W	108W

TABLE 3.4.1

AD-A172 362 OPTICAL PROCESSING(U) DAYTON UNIV OH RESEARCH INST 2/2
A C WALKER ET AL. JUN 86 N00014-85-K-0479

UNCLASSIFIED F/G 20/6 NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

HOE Specification

For Dichromated Gelatin Holograms

	Present Capability				Target Specification	
Operating Wavelength λ	514nm				850nm	
Continuous Irradiance (1), (1A)	$100 W/cm^2$				$100 W/cm^2$	
Operating Temperature	0-70°C				0-70°C	
Resolution (1)	5,000 lines/mm				5000 lines/mm	
Refractive Index Modulation (2)	0.08				0.08	
Diffraction Efficiency (3)	95% (Minimum)				98% (Minimum)	
Exposure Sensitivity of Dichromated Gelatin	UV and Blue Green				UV and Blue Green	
Practical Thickness (Maximum)	100μm				100μm	
Minimum Thickness Producing Diffraction Efficiency > 90%	5μm				5μm	
HOLOGRAM MEDIUM	THIN		THICK		THIN	
MODE OF DIFFRACTION	TRANSMISSION		TRANSMISSION		TRANSMISSION	
PROPERTY MODULATED	Amplitude	Phase Shift	Absorption Const		Absorption Const	
Maximum Theorectical Efficiency (4)	6.25	33.9	3.7	100	7.2	100
Maximum Efficiency Obtained Experimentally	6.0	32.6	3.0	90	3.8	80

Notes

(1) Ref. R.A. Cullen SPIE Vol 369 Max Born.

(1A) Processed layer does not darken or distort on intense exposure in the wavelength range 430-520nm. In the absence of an absorption bound in gelatin at longer wavelengths, then this damage threshold will apply out to 1550nm.

(2) Measured for gelatin layers up to 50μ thick.

(3) Diffraction Efficiency $= \frac{P.0 \text{ Order}}{P \text{ illumination}}$ where P = power

(4) Hologram Efficiency = n

AND n = SEoV where S = Holographic Sensitivity
Eo = Average Exposure
V = Fringe Visibility

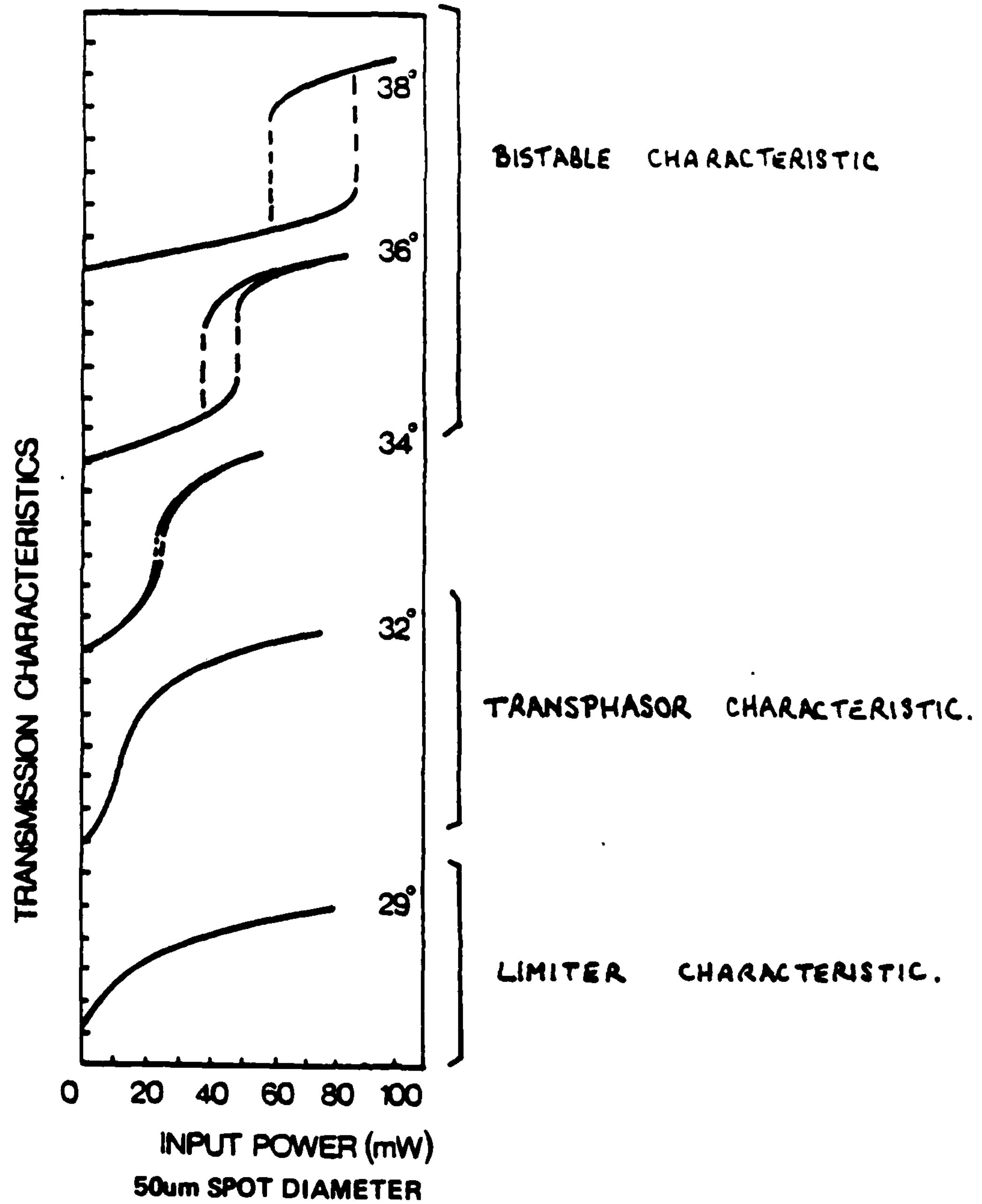
TABLE 3.4.3.3

SLM Requirement

	Present Performance	Target Specification
Number of Pixels	5.1×10^3	86×10^3
Area of Spatial Light Modulator	1 cm^2	1 cm^2
Resolution No pixels/mm of Etalon	10 pixels/mm	50 pixels/mm
Frequency of Operation		50 KHz
Write Time (MAX)	-	10μs
Erase Time (MAX)	-	10μs
Storage Time	1 hour	1 hour

TABLE 3.4.3.5

21585



Variation of Transmission Characteristics with Initial Detuning for a ZnSe Interferometer	Figure 3.1

21586

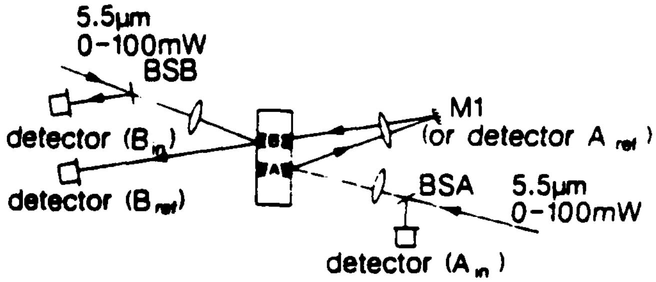


Figure 3.2.A

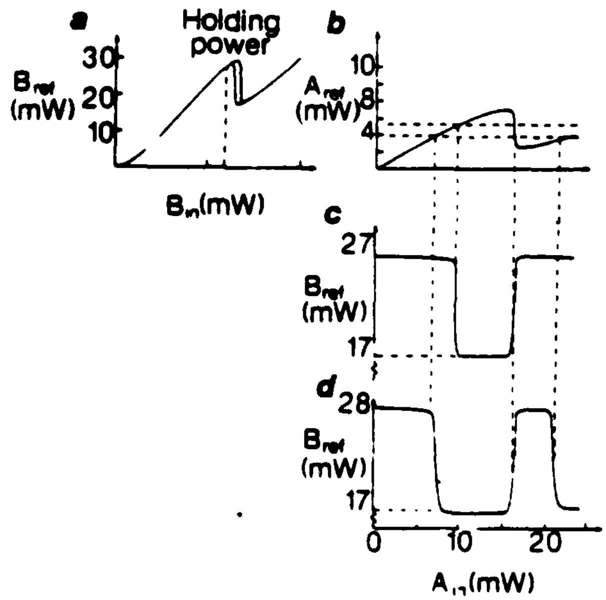
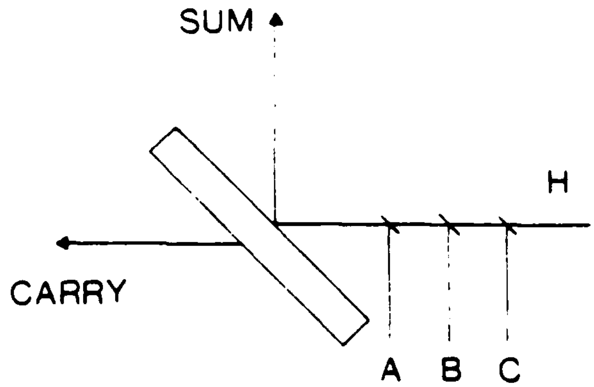
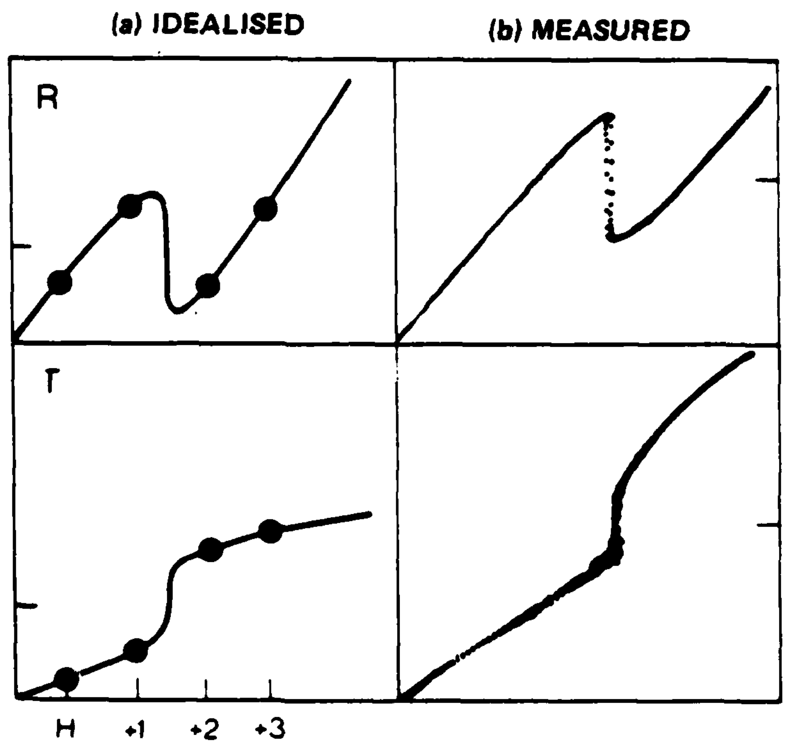


Figure 3.2.B

Exclusive or Implementation	Figure 3.2.A & B

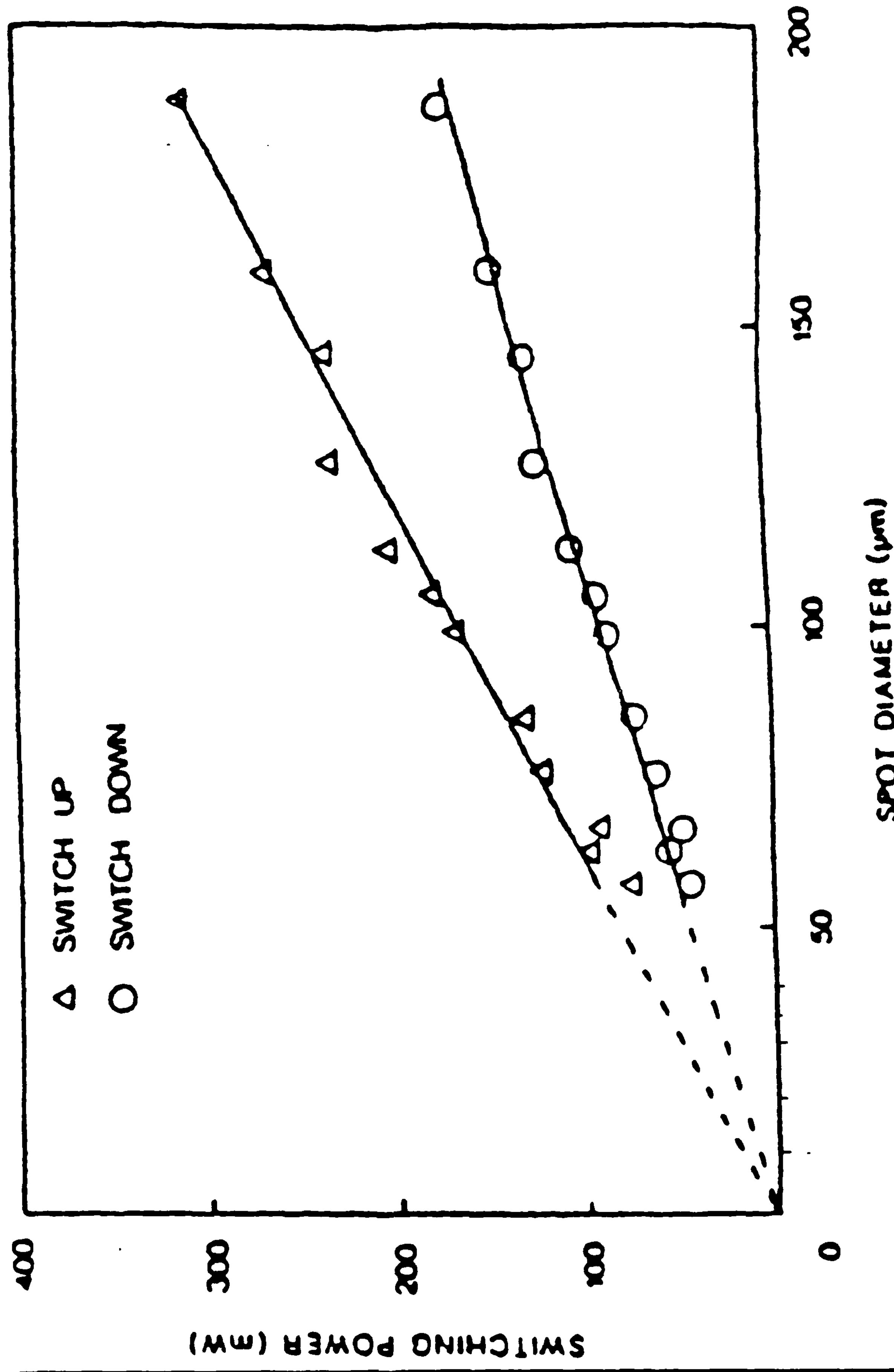
21587





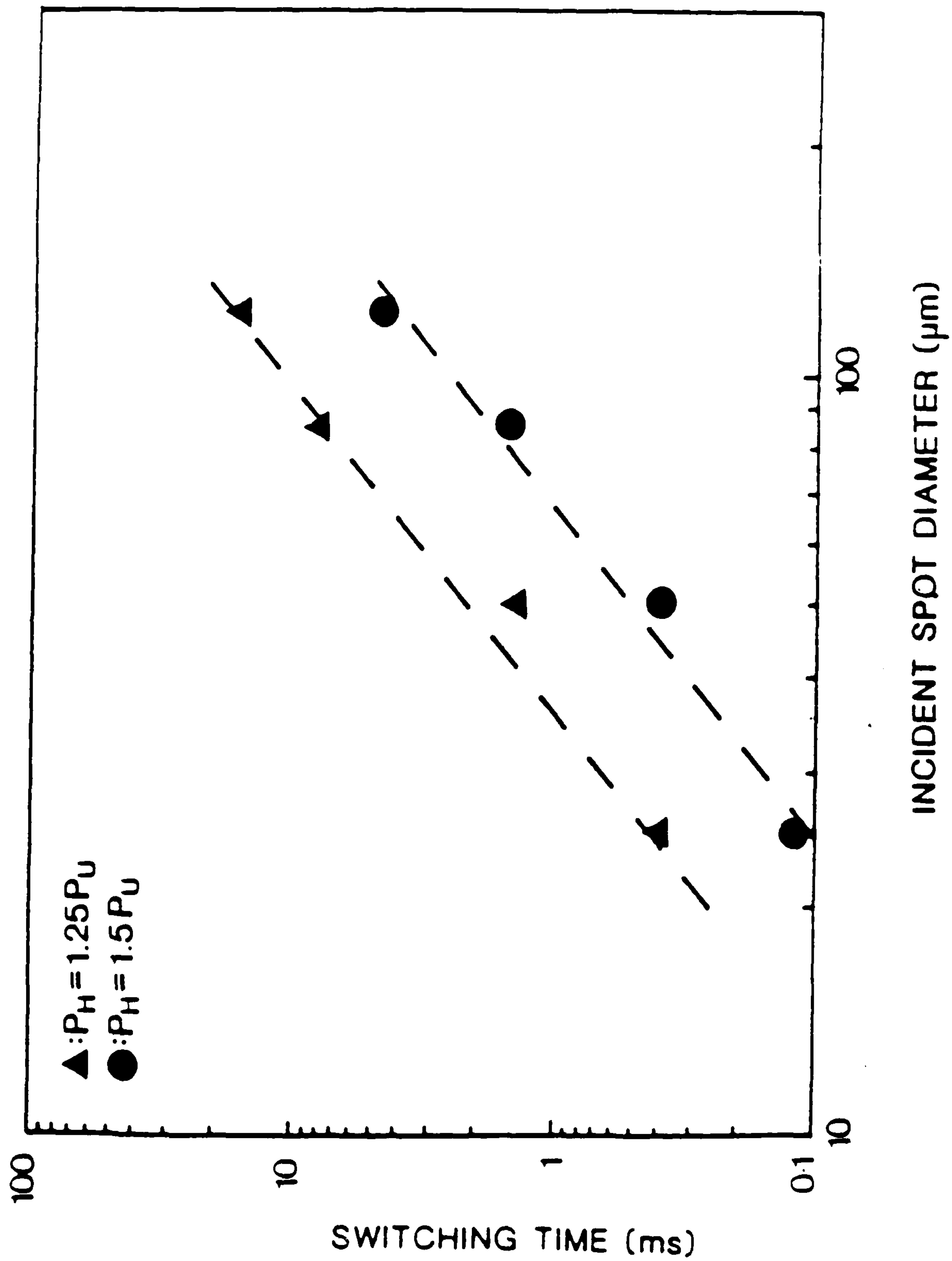
Schematic address and Output of a Full-Adder Plate	Figure 3.3

21588



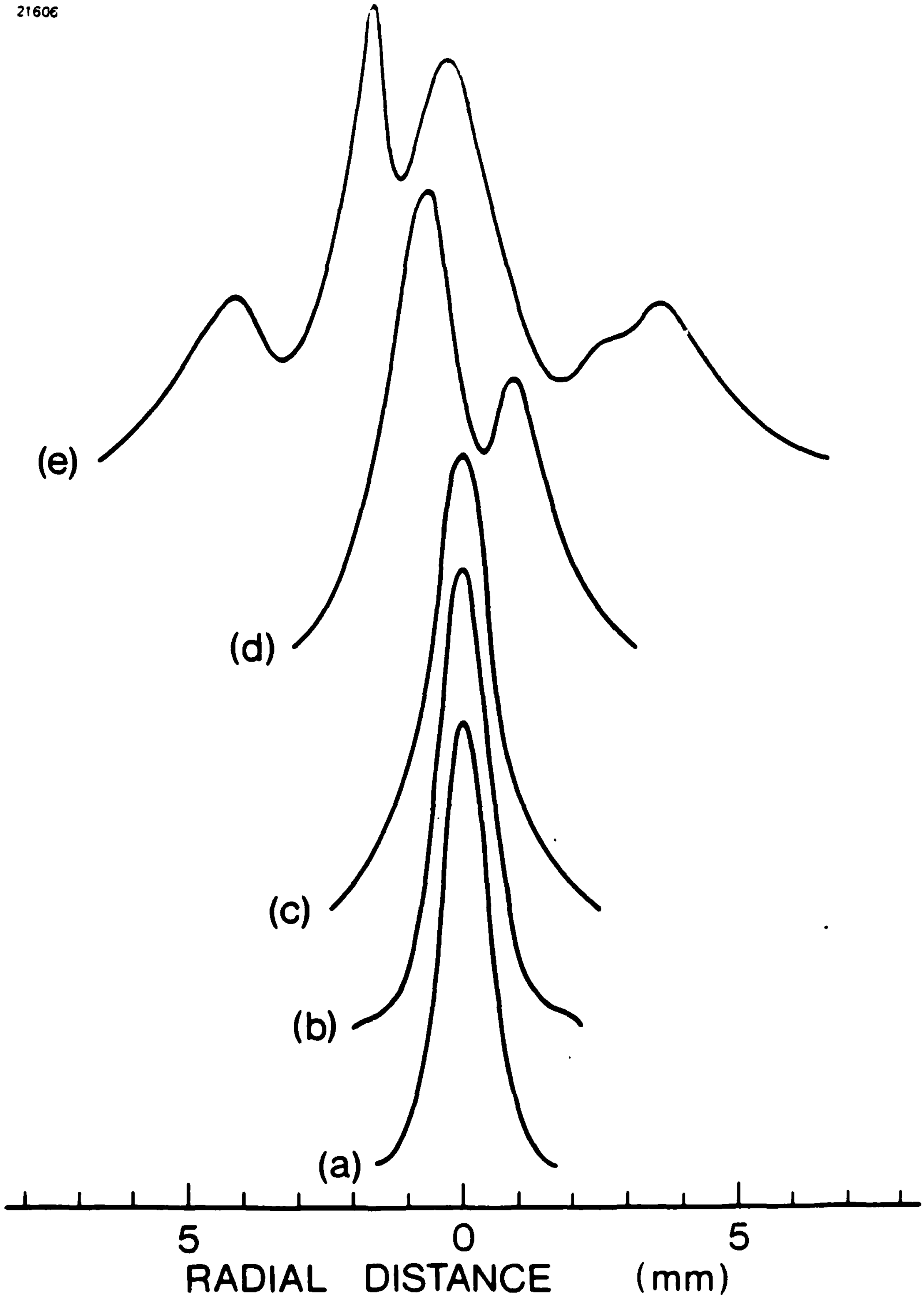
Spot Diameter Dependence of Switching Powers	Figure 3.4

21591



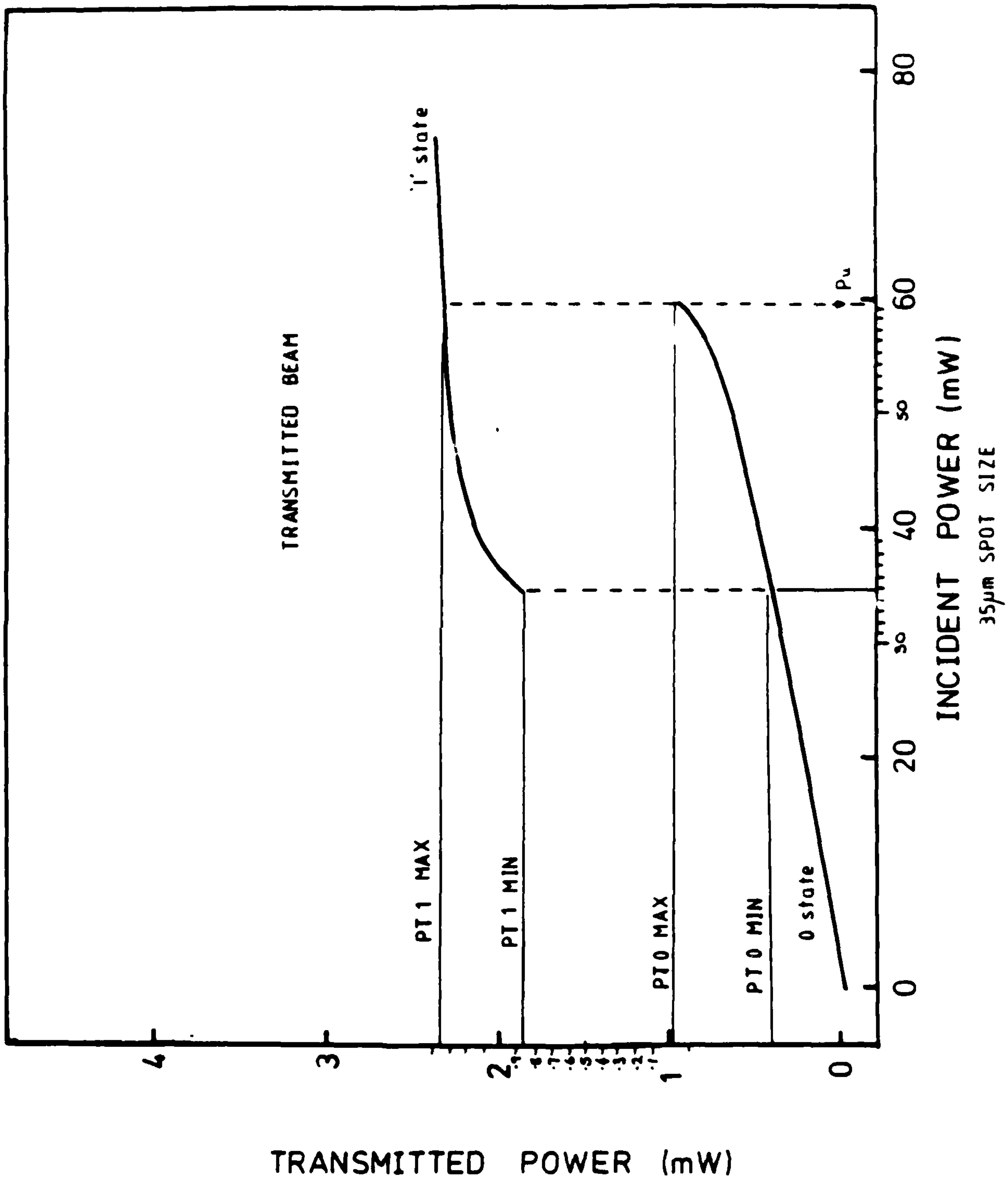
Variation of Switching Time with Spot Diameter	Figure 3.5

21606

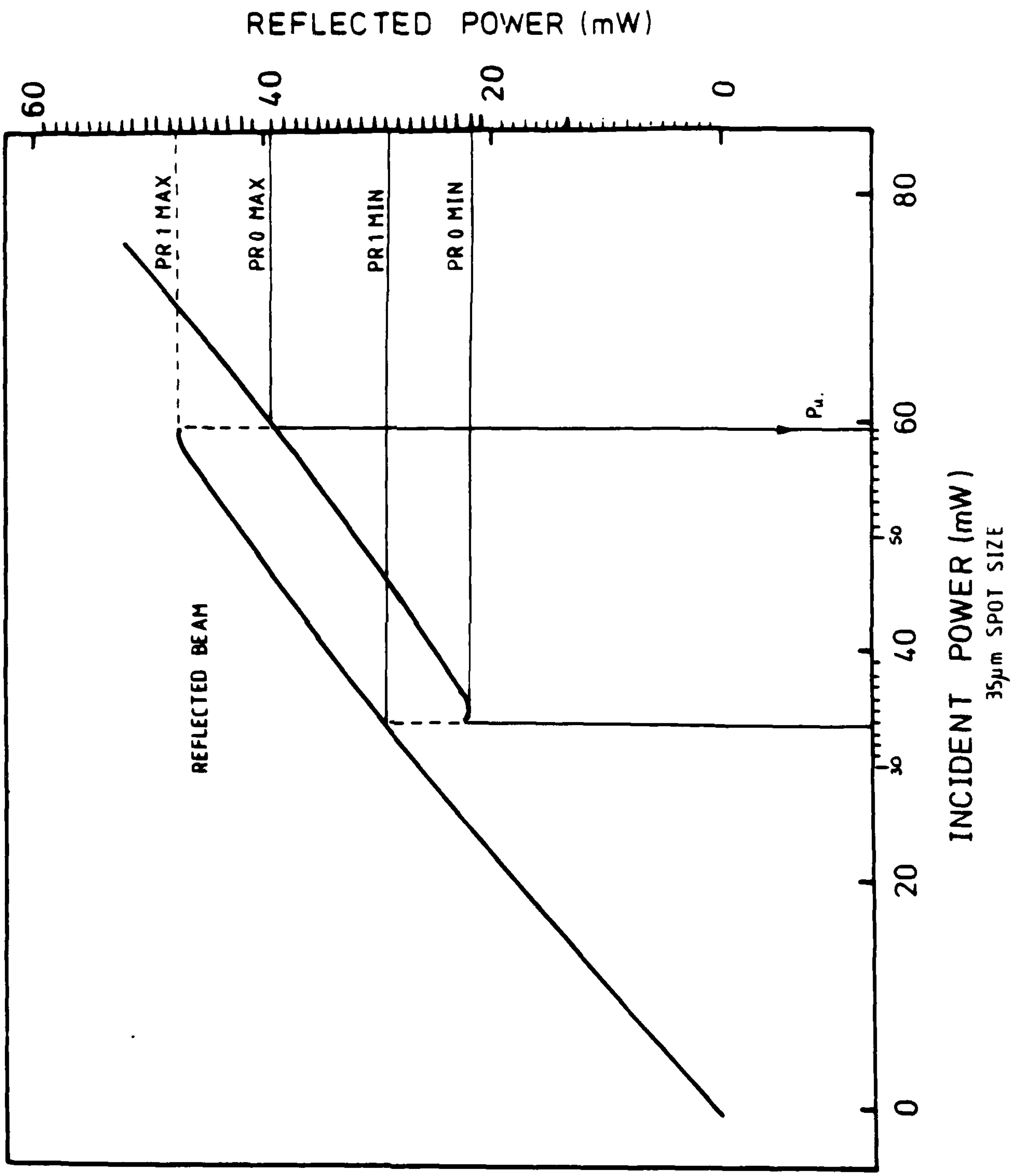


(a) 25mW; (b) 33mW; (c) 50mW; (d) 66mW; (e) 100mW
BEAM PROFILE MONITORED 10cm BEHIND SAMPLE

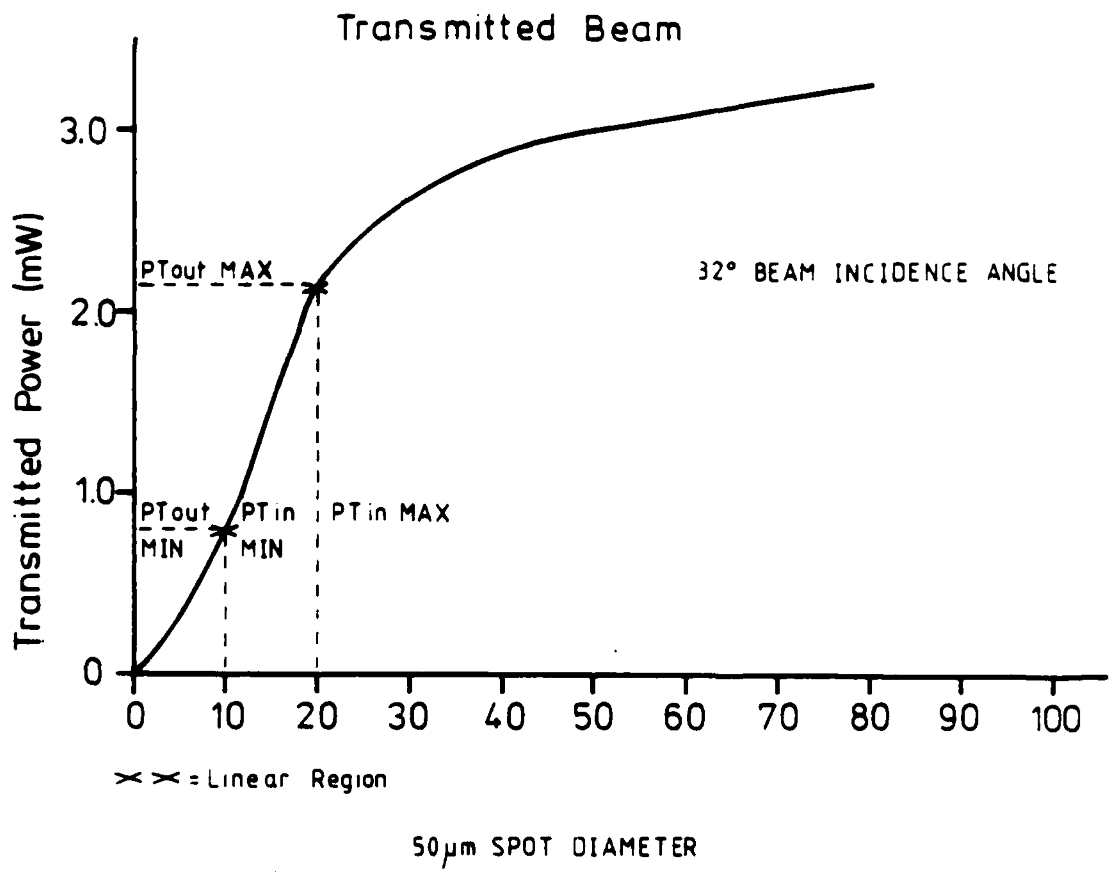
Beam Profile of a Transmitted Beam	Figure 3.6



Typical ZnSe Bistable Interferometer Transmission Characteristic	Figure 3.7



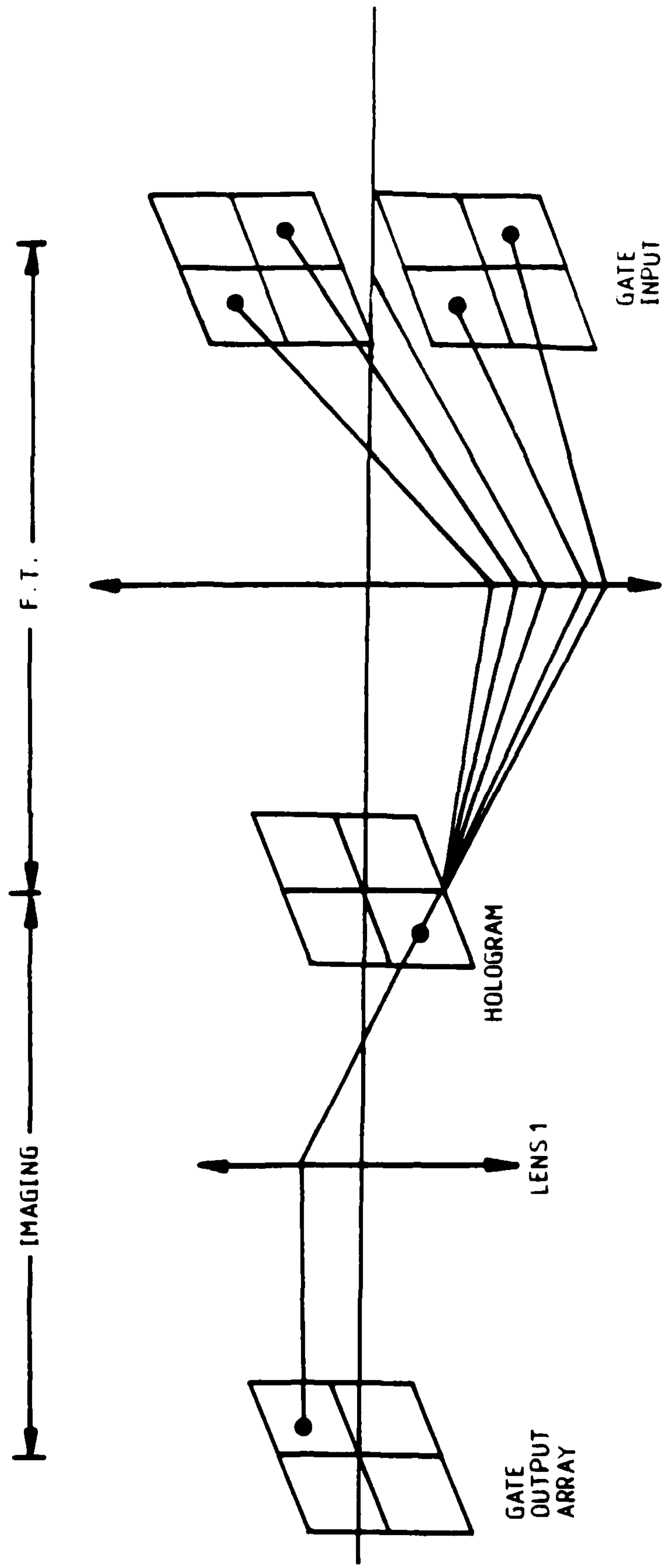
Typical ZnSe Bistable Interferometer Reflection Characteristic	Figure 3.8



MEASURED TRANSPHASOR CHARACTERISTIC.

Measuring Transphasor Characteristic	Figure 3.9

21590



Optical System for the Direct Implementation of Space-Variant Interconnections	Figure 3.10

4. BERLEKAMP - MASSEY ALGORITHM

Error-control codes are concerned with techniques for the protection of digital data against errors that occur during transmission. One important class of such codes is the Bose-Chaudhuri-Hocquenghem (BCH) class of multiple-error-correcting codes (of which the Reed-Solomon codes are a well-known sub-class).

The Berlekamp-Massey algorithm is an efficient technique which may form part of the procedure for decoding BCH codes, but also has more general application in the autoregressive filtering of data. Massey [MASS69] has shown that the algorithm may be viewed as a procedure for designing a minimal-length linear feedback shift register (LFSR) to generate a given finite-length sequence of digits.

A significant proportion of the computation required to decode BCH codes involves the solution of a matrix equation of the form:

$$\begin{bmatrix} S_1 & S_2 & S_3 & \cdots & S_v \\ S_2 & S_3 & S_4 & \cdots & S_{v+1} \\ S_3 & S_4 & S_5 & \cdots & S_{v+2} \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ S_v & S_{v+1} & S_{v+2} & \cdots & S_{2v-1} \end{bmatrix} \begin{bmatrix} C_v \\ C_{v-1} \\ C_{v-2} \\ \cdot \\ \cdot \\ \cdot \\ C_1 \end{bmatrix} = \begin{bmatrix} -S_{v+1} \\ -S_{v+2} \\ -S_{v+3} \\ \cdot \\ \cdot \\ \cdot \\ -S_{2v} \end{bmatrix}$$

for vector C.

This is an equivalent problem to the synthesis of an autoregressive filter for the sequence S_j :

$$S_j = - \sum_{i=1}^{V} C_i S_{j-i} \quad , \quad j = V+1, \ldots, 2V$$

where the C_i (or C_v above) may represent the coefficients of the error-locator polynomial in the context of decoding BCH codes. The C_i also represent the "weights" of the feedback taps in a minimal-length LFSR which generates the sequence S_j (see figure 2.5.1).

Berlekamp [BERL68] discovered an iterative algorithm for finding the error-locator polynomial when decoding BCH codes. Massey [MASS69] reformulated the algorithm as a procedure for designing minimal-length LFSR's, and described a logical circuit for implementing the algorithm in hardware.

4.1 INVESTIGATION REQUIREMENTS

In considering the Berlekamp-Massey (B-M) algorithm, we were primarily motivated by suggestions from the University of Dayton that such an investigation would be complementary to the work of other members of the optical computing community. Our goal then, was to identify a suitable optical realisation of the B-M algorithm, based upon optically-bistable etalons implementing digital logic functions. In order to simplify the physical requirements, we restricted our consideration to binary-valued input data sequences. (NOTE: not simply binary representations of multiple-valued input data sequences). This restriction allows us to use a minimal number of logic gates and to perform modulo 2 arithmetic.

4.2 OVERVIEW OF THE BERLEKAMP-MASSEY ALGORITHM

We can view the purpose of the B-M algorithm as the synthesis of a minimal-length LFSR of the form shown in Figure 4.1. The parameters of the LFSR that we seek to obtain are the "weights" associated with the feedback taps, and the length (number of stages) of the LFSR.

Briefly, the synthesis proceeds (with reference to Figures 4.2 and 4.3) as follows. Initially, we assume a LFSR of zero length, and evaluate the resulting error in generating the first syndrome of the sequence (S_o). If there is no error then the current LFSR design is correct, but if an error is present, the error value is used to determine the appropriate feedback tap(s) for an LFSR which would generate the correct sequence. This process of modifying the current LFSR design in the presence of an error involves changing the LFSR feedback tap weightings and sometimes increasing the length of the LFSR as well. In the next iteration the procedure is repeated, this time evaluating the error in generating the sequence (S_o,S_1) using the LFSR designed in the first iteration. In subsequent iterations the error in generating the sequence (S_o, S_1, S_2, ...) is evaluated using the LFSR designed during the previous iteration, until the final iteration for the sequence (S_{E-2}, S_{E-1}, S_E, S_{E+1}, ... S_{2E-1}) resulting in (C_o, C_1, C_2, ... C_L) feedback tap weightings where L is the designed length of the LFSR. For a more detailed description of the algorithm see [BERL68], [MASS69], [BLAH83] and [L1U84].

4.3 DESIGN CONSIDERATION

The technique of data flow analysis was originally employed by designers of optimising compilers for high-level programming languages. This technique later had a great influence on many computer designers, resulting in a considerable amount of research into data flow architectures [DENN80, TREL82, VEGD84] of which the systolic array concept [KUNG79] can be considered a special case. By analysing the flow of data between instructions and statements in a program one can derive a data

flow graph. Such a graph represents the dependencies of each instruction or statement, on the prior availability of input operand data and a destination for the result, in the form of a digraph. This has the benefit of exposing the maximum amount of parallelism available in a particular algorithm or program, and forms the basis of a program for a data flow computer.

A fairly superficial data flow analysis of the B-M algorithm quickly reveals that the procedure cannot be linearised to allow the concurrent evaluation of each iteration. This is due to the data dependency of successive evaluations of d_n (the discrepancy or error in generating the current sequence) upon the values of C_1, C_2, C_3, ... (the feedback tap weightings) from the previous iteration. Thus, we may conclude that there is only very limited scope for parallelism, and that maximum performance will be achieved through the use of pipelined operations. This is not to suggest that a different algorithm might not be designed with greater inherent parallelism, but this is beyond the scope of our investigations.

4.4 LOGICAL STRUCTURE

Considering the limitations of the B-M algorithm discussed in the previous section, we decided to adopt a systolic approach in our design. Such an approach has a number of benefits including:

- implementation as a regular array of identical cells
- regular flow of control and data
- maximised throughput using pipelined operations
- linear relationship between the number of cells and the length of the input sequence.

Also, this approach is well-suited to an implementation using optically-bistable etalon devices due to the simplicity of each cell.

Our realisation of the LFSR synthesis circuit owes much to the original work of Massey [MASS69] and the VLSI implementation suggested in [L1U84]. However, we believe that the systolic cell organisation of the circuit proposed by us, is a further useful development for either VLSI or optical implementation.

In designing the systolic cell for the special case of binary-valued syndromes we anticipated some reduction in complexity. This is so, however it turns out that the complexity can be reduced even further by making the following observations (refer to Figures 4.2 and 4.3):

(a) there are essentially 3 cases in the algorithm -

(i) $d_n = 0$

(ii) $d_n \neq 0$ and $n >= 2L$

(iii) $d_n \neq 0$ and $n < 2L$

(b) d^* is initialised to the value 1 and is only modified for case (ii) where $d_n \neq 0$ and $n >= 2L$.

(c) for binary-valued syndromes, if $d_n \neq 0$ then $d_n = 1$ and hence:

$$\underline{d^* = 1 \quad \text{(always)}}$$

Furthermore, since $-d_n/d^*$ is only evaluated when $d_n \neq 0$, then:

$$\underline{-\frac{d_n}{d^*} = 1 \text{ (always)}}$$

Hence, $-d_n/d^*$ need never be evaluated and also any multiplication by $-d_n/d^*$ is redundant since the result will be unaffected.

With reference to Figure 4.2 all multiplication operations shown in the "Upper Logic" can be deleted, and the evaluation of $-(d^*)^{-1}$ is completely unnecessary. These simplifications result in the systolic cell shown in Figure 4.4 which is interconnected to form a linear array as shown in Figure 4.5.

4.5 OPTICAL IMPLEMENTATION

As mentioned earlier, our implementation is based upon non-linear interferometers such as those developed by Heriot-Watt University, Edinburgh, providing the primary switching characteristics.

Each cell of the systolic array comprises ten interferometers, two beamsplitters and one mirror. Figure 4.6 illustrates the interconnections and suggests a possible geometry assuming each interferometer has a fan in of 2 and a fan out of 2 capability. The holding beams and any necessary focusing elements have been omitted for clarity. Though adhering strictly to the systolic philosophy, we recognise that the implementation we propose could be more efficient in time if the summation to produce d_n was carried out in parallel rather than passing results left to right from cell to cell. The operation of a single cell is as follows:

Step 1

All syndromes are shifted one place to the left, hence a new bit is stored in memory S_n. At the same time all the B_i are shifted one place to the left (with "n >= 2L" forced to be false/zero at this time), hence B_i takes the value of B_{i-1}.

Step 2

Using De Morgan's rule, an OR function is used to generate the modulo-2 product of S_n and C_i (which are inverted due to the nature of beams reflected from the non-linear interferometers). This product is added (modulo-2, using an EXOR function) to the partial sum generated in the cell immediately to the left. The result is another partial sum which is presented to the cell on the immediate right. The total of this summation process appears as the PARTIAL SUM OUT in the right most cell of the array, and represents the value of d_n (see Figure 4.5).

Step 3

If the value of d_n is 0, the n - counter (not yet implemented optically) is incremented, and the process continues from Step 1 unless n = 2E-1.

If the value of d_n is 1 and n >= 2L:-

(a) B_i is overwritten with the old value of C_i,

(b) C_i is overwritten with the sum of the old value of C_i and the old value of B_i,

(c) All the B_i are shifted one place left (with a '1' loaded into the left most cell),

(d) The value of the LFSR length is modified such that: L := n + 1 - L (not yet implemented optically).

If the value of d_n is 1 and n < 2L :-

(a) B_i retains the same value,

(b) C_i is overwritten with the sum of the old value of C_i and the value of B_i,

(c) All B_i are shifted one place to the left.

Step 4

If $n = 2E - 1$, the calculation is complete with the results held in memory C_i of each cell, the first L of which are valid. If $n < 2E-1$, then increment the n - counter and return to Step 1.

The strength of this optical implementation lies with the capability of a single cell array to provide all the systolic cells necessary to decode any length of sequence up to 2×10^4 syndromes assuming a 10^4 capability of each array of interferometers. The output signals of one systolic cell can be fed back to become the input signals of another systolic cell, with each systolic cell defined by its position on the interferometer arrays.

Additional optical control is required to ensure that the input of any systolic cell does not track the input of the previous systolic cell. This additional control may be provided by a "lock and clock" technique, similar to that described by Walker [WALK86], as shown in Figure 4.7. Note that all of the systolic cell outputs are already registered, thereby reducing the requirement to one external memory cell from the two shown in Walker's implementation.

Control of the memory elements may be achieved through the use of acousto-optic modulators controlled themselves by a microprocessor. As previously stated, the strength of the implementation does not lie with its processing rate but with its flexibility to decode a sequence of binary-valued syndromes of any length up to 10^4. It is not expected that the microprocessor control will limit the processing rate of the system, based on current ZnSe interferometer switching rates. Should the switching rates increase however, then further all optical control may be required.

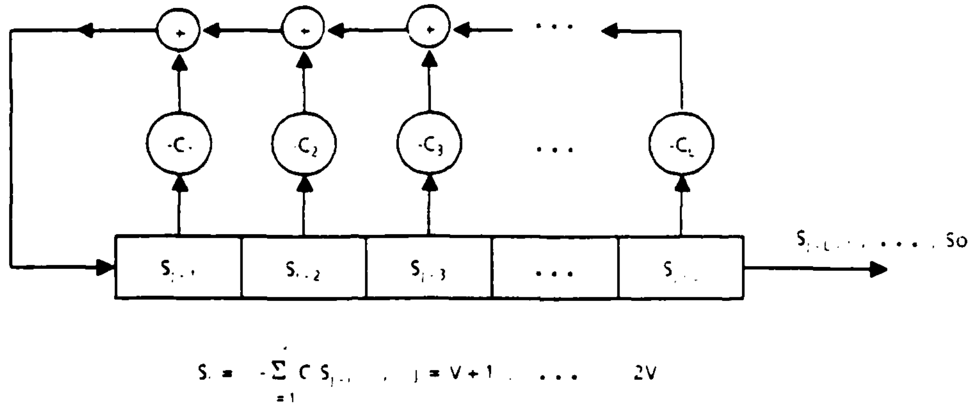


Figure 4.1 General L-Stage Linear Feedback Shift-Register (LFSR)

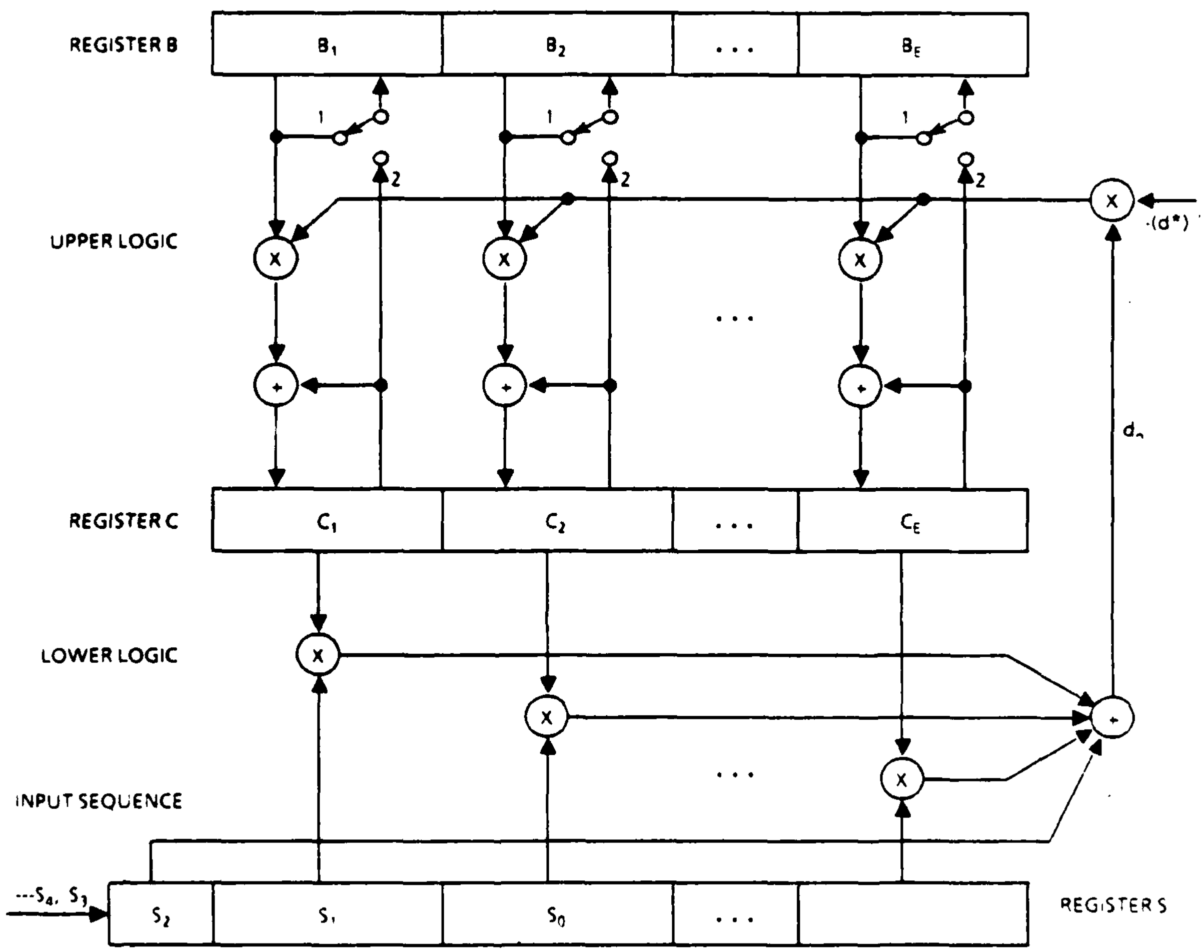
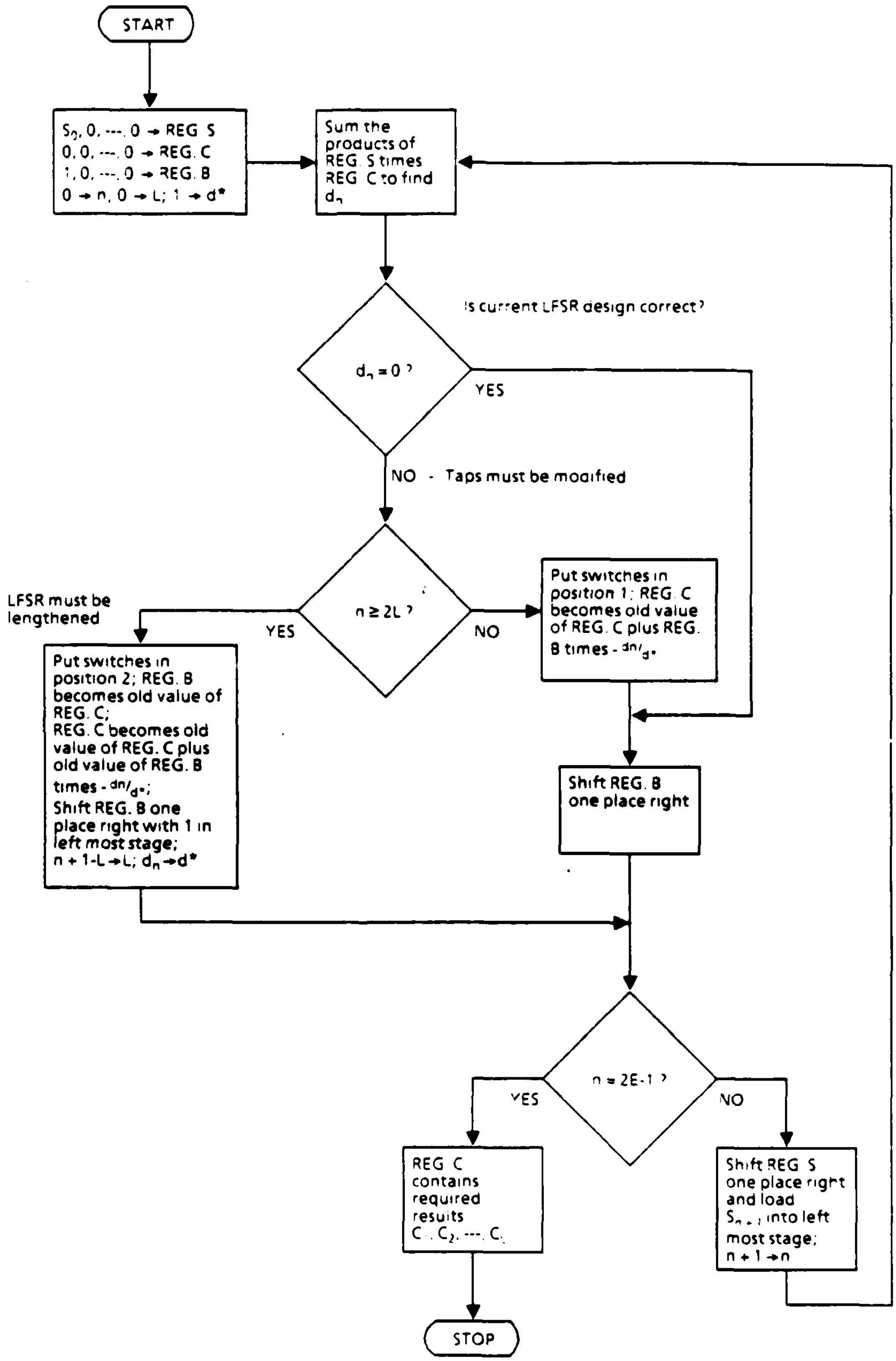


Figure 4.2 Berlekamp-Massey LFSR Synthesis Circuit



Berlekamp-Massey LFSR Synthesis Algorithm	Figure 4.3

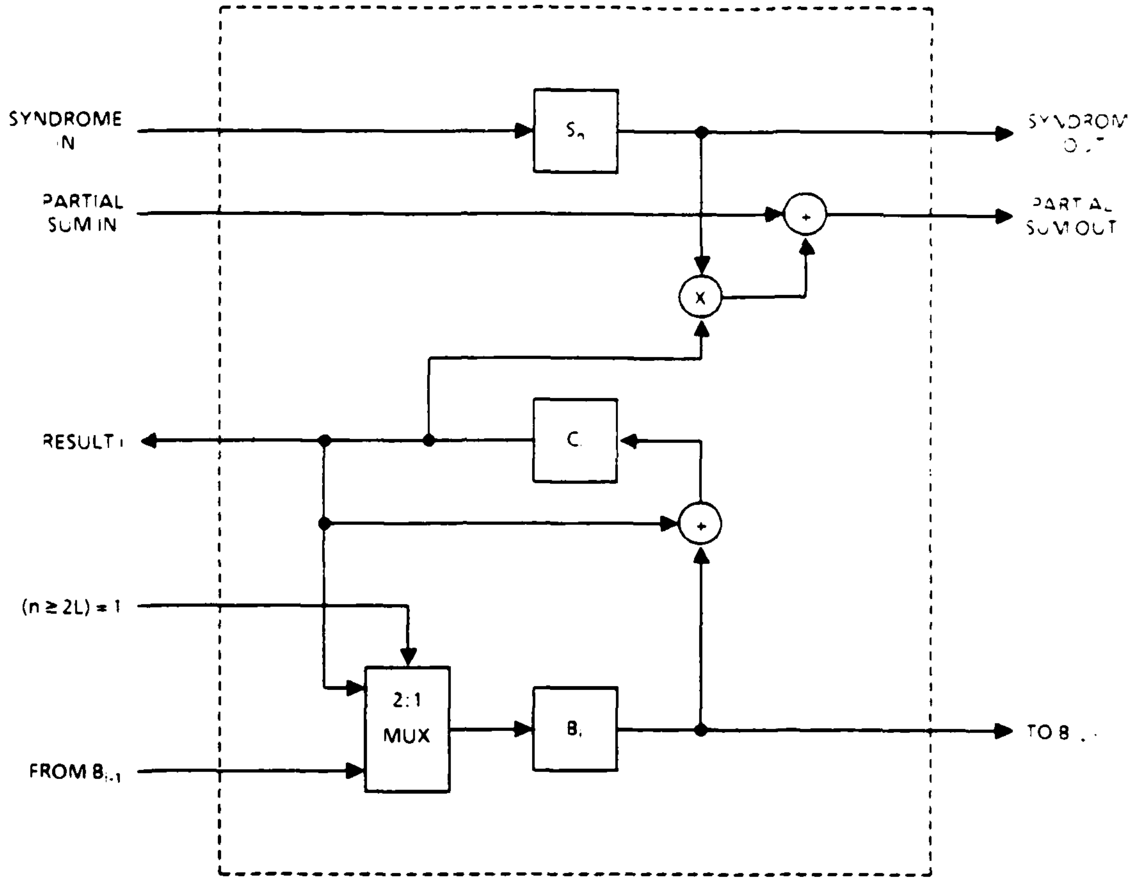
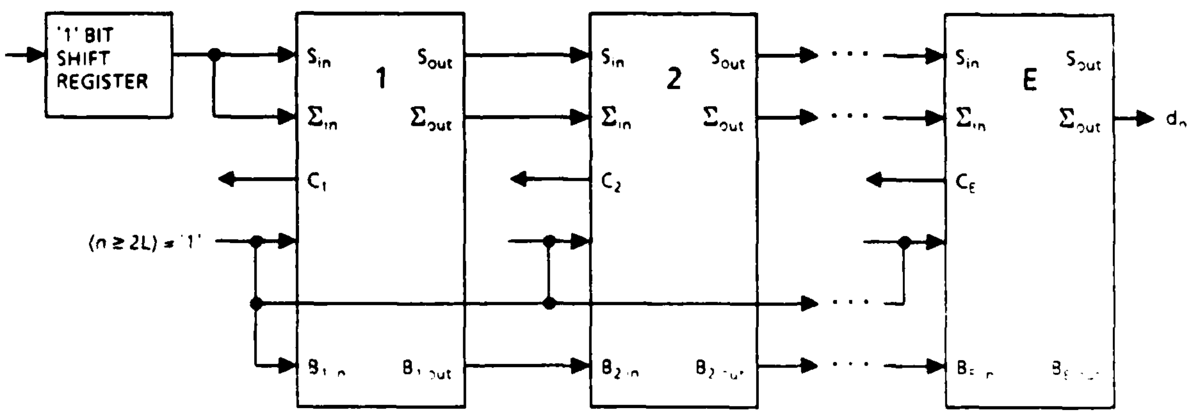
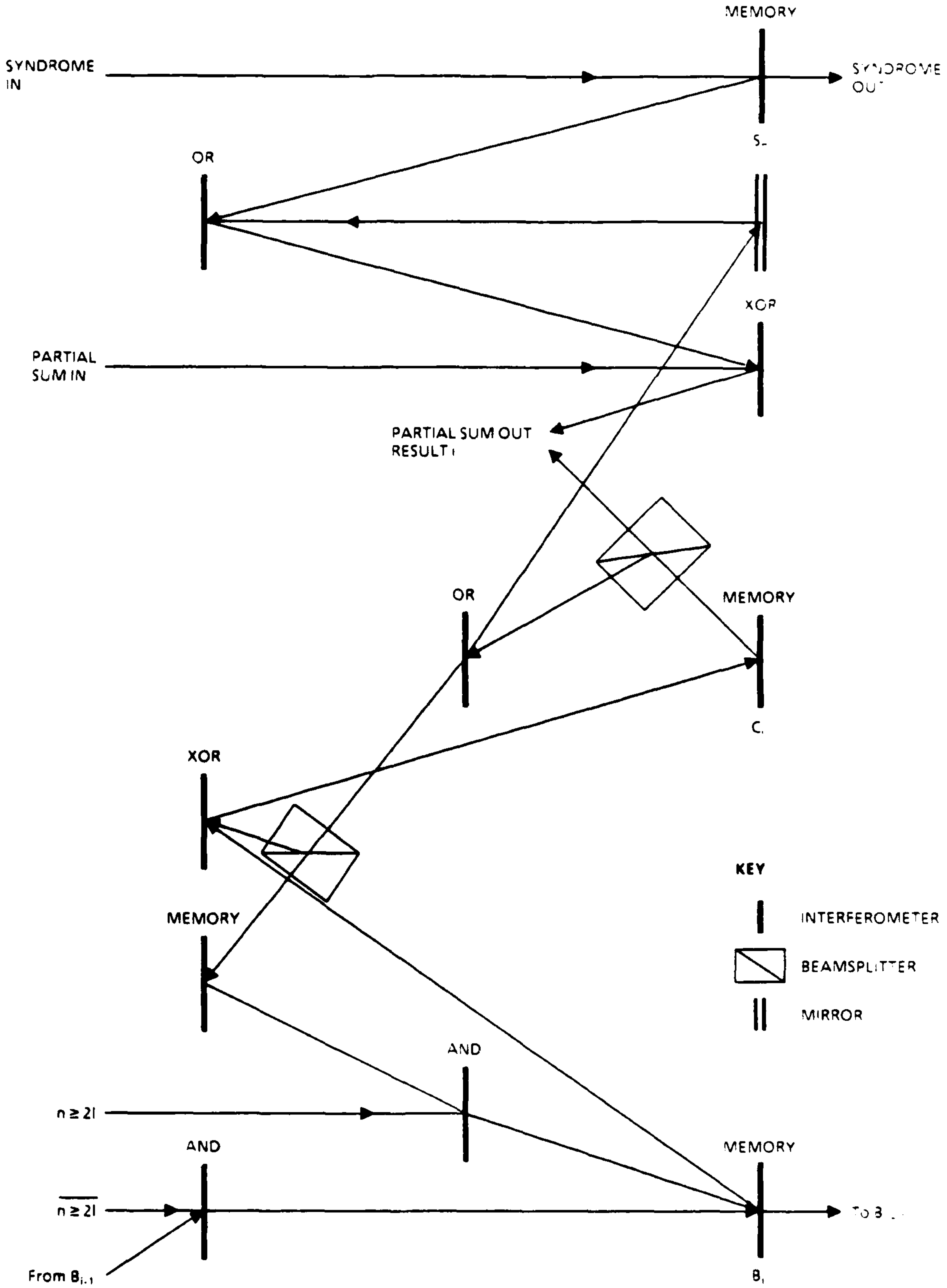


Figure 4.4 Logical Function of Systolic Cell for Binary-Valued Syndromes

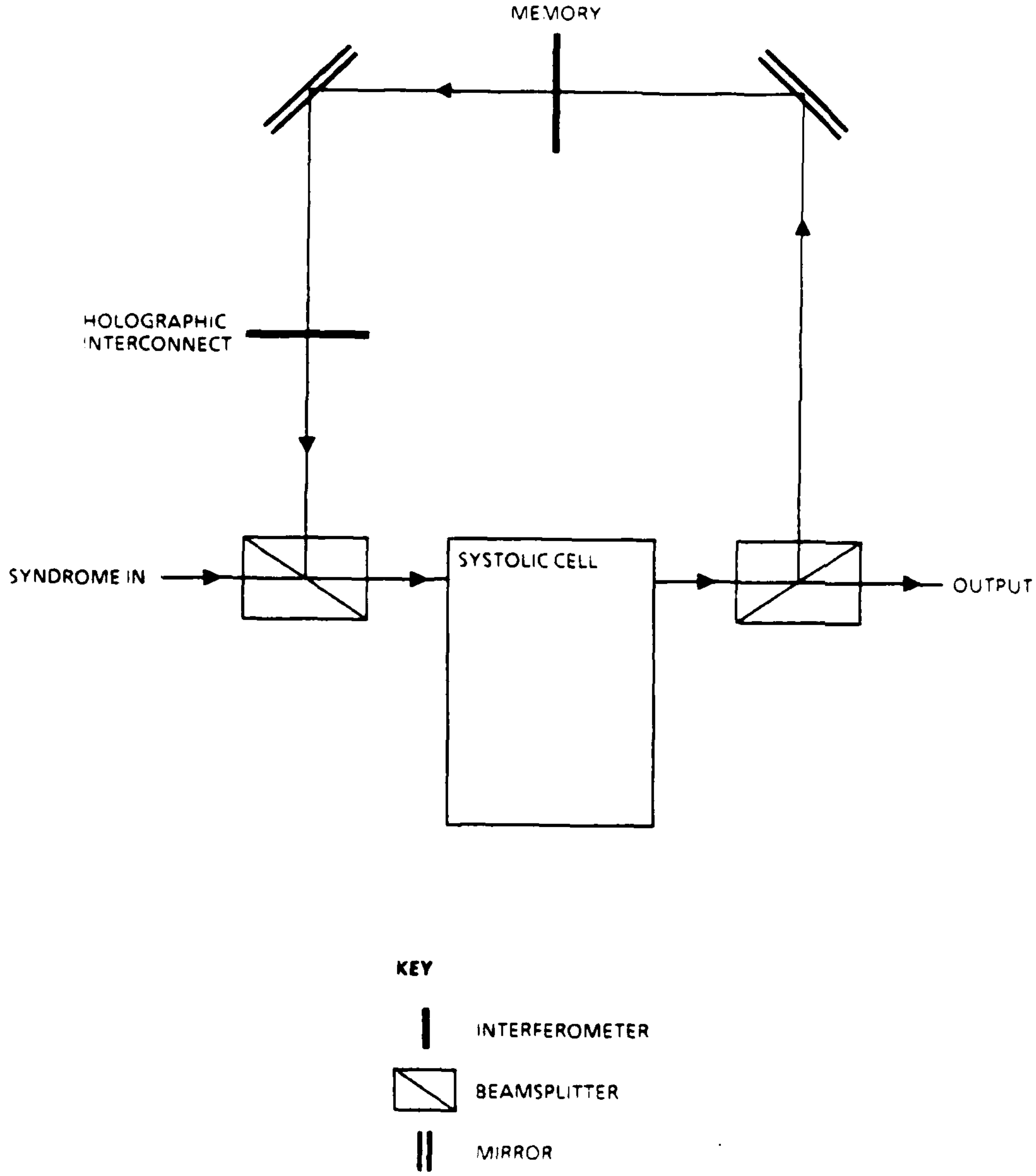


Not shown
1 Control logic for sequencing within each cell
2 Evaluation of [($d_n \neq 0$) AND ($n \geq 2L$)]
3 Evaluation of [$L := n + 1 - L$] and [$n := n + 1$]

Figure 4.5 Schematic Showing Interconnection of E Systolic Cells for an Input Sequence of 2E Syndromes



Optical Implementation of Systolic B-M Cell for Binary Valued Syndrome	Figures 4.6



Peripheral Control of B-M Systolic Cell Containing Registered Outputs	Figure 4.7

5. GENERAL REVIEW OF ACHIEVEMENTS

The major objectives of the initial phase of the program have been achieved and significant considerations have been highlighted and discussed, with recommendations for future areas of investigation and development.

The main observations are detailed below.

5.1 APPLICATIONS AND ARCHITECTURES

For applications demanding high performance using complex systems employing optical and opto-electronic device implementation the following list of general requirements is identified:

(a) Modularity

(b) Standardisation

(c) Fault tolerance

(d) Availability

(e) Correctness.

These attributes can only be supported by suitable hardware design and software structure, and cannot be introduced retrospectively.

No single architecture is optimally suited to all the applications considered. Five general architectural styles have been classified and appraised. These are identified as Systolic arrays, Supercomputers, MIMD Computers, Data Flow and Reduction Computers, and Massively Parallel Computers (which incorporate Neural Networks and Connection Machines).

The assessment concludes that systolic arrays appear best suited for very high performance fixed and regular functions, especially "computationally intensive" functions. In a basic form, developed for VLSI, this architectural style is not easily modified to perform enhanced functions and may not easily facilitate error containment and recovery mechanisms essential for fault tolerance. However, the introduction of hardware redundancy, leading to fault-tolerance capability, is envisaged by exploiting the potential availability of global interconnectivity that optical technology provides.

The Data Flow architectural style appears well suited for many high performance and complex general purpose applications, where the processing can be described as "data-driven" or "demand-drive". An important benefit of this architecture is that it supports the use of a Functional Language. Parallel processing is implicit when written in this language (i.e. Parallel processes do not need to be explicitly defined within the written program), and the processing response derived by the architecture is mathematically equivalent to the program statement. Hence program generation and verification are significantly improved by comparison to other languages and architectures.

The fine-grain parallelism found in Data Flow models is well suited to optical implementation, although development and research is required to establish the simpler processing element structures and dynamic global interconnection techniques believed essential for successful implementation of the "static" model. Further techniques need investigation to perform data token matching, required to implement the "dynamic" model successfully.

Massively Parallel Computers refer to those architectures requiring a very large number of processing nodes which are best suited for nearest-neighbour search operations fundamental to the tasks of pattern recognition, associative memory and error correction. These are typical of those functions which demand processing capabilities of many orders

greater than conventional electronic computer techniques. The study concludes that the essential requirements for this type of architecture are for large numbers of parallel processing elements, each of which performs a relatively simple function. In addition, there is a need for both global and dynamic interconnections between processing elements. Consequently this architecture type is likely to gain substantial benefits from optical techniques for implementation.

Massively Parallel Computers appear to be able to exploit the unique properties of an optical implementation, to yield a substantial increase in processing capability. Successful implementation will depend on further development and research into the hardware requirements for ultra-simple processing elements and dynamic global interconnection techniques.

The remaining architectural styles offer less flexible solutions with attendant complications in design and development.

5.2 TECHNOLOGY

A comprehensive study and assessment has been performed with the main emphasis applied to the use of optical non-linear interferometers, and the identification and use of compatible supporting technologies for digital or discrete number functions.

An important outcome of the program was to devise detailed specification formats for the devices under consideration. These are essential if all performance criteria are to be met in the research into improved and optimised characteristics, and equally important if any meaningful assessments of comparative performances are to be made.

Specifications were completed for current performance and projected requirements, based upon an initial estimation of requirements thought likely for the preferred architectures. Scrutiny of these specifications identify the following important observations:

(a) There is a requirement for etalon based interferometers to operate at CW at near infra-red wavelengths, so that solid state laser diode power sources and fiber interconnection could be used with efficiency.

(b) ZnSe Interferometers must be developed to consume substantially less power than is presently attainable.

(c) It is anticipated that a reduction in spot size, together with the development of a physical construction of the etalon which maintains acceptably low power dissipation and cross talk, will reduce incident power levels, reduce switching times and achieve a significant expansion of processing capacity due to increased spot density, for ZnSe non-linear interferometers.

(d) Solid state laser diodes should be used as power sources because they are more compact, more efficient and easier to use than other laser sources. Further advances are required to produce solid state laser diodes which will perform at 514nm wavelengths in CW and with sufficient output power.

(e) Detector technology appears to be sufficiently advanced to meet the forecast requirements. However, larger arrays will need to be developed in order to be compatible with projected array sizes.

(f) Holographic Lens appear to be most suitable for power (holding) beam array generation and alternative static interconnection requirements.

(g) Suitable techniques and devices must be developed which can dynamically switch and reroute large arrays of light. A great deal of research activity is devoted to Spatial Light Modulators, which have potential in this important area.

A specific analysis of two input OR/NOR gate and memory configuration, proposed to illustrate a minimum general purpose requirement, was carried out to evaluate performance criteria for ZnSe non-linear interferometers. Expressions have been derived to quantify minimum values for Power Transfer Rates and Contrast Ratio, which can be used directly to assess the suitability of interferometer performance characteristics to enable them to operate in cascade. These expressions also identify tradeoffs in performance, and hence are extremely useful when determining optimised characteristics for future research and developement activities.

An appraisal of published ZnSe Bistable Interferometer characteristics reveals that these devices do not perform adequately in the general purpose configuration either in transmission or reflection mode of operation. The minimum requirements for correct operation have been stated. To summarise, it is essential for the Power Transfer Ratio for transmission to be improved, and for reflection this ratio may be reduced. It is also essential to improve the Contrast Ratio in reflection.

A number of methods and approaches for achieving these improvements have been proposed.

The structural requirements for implementation have been discussed and suggestions are made for possible device types which would be advantageous in a practical construction.

5.3 BERLEKAMP-MASSEY ALGORITHM DESIGN STUDY

The Berkelamp Massey Algorithm was adopted for a design study of an optical implementation, based on non-linear interferometers of the type developed by Heriot-Watt University. The study dmonstrates the use of Data Flow Analysis, which is able to discriminate parallel processes in a structured method of design.

As the algorithm was conceived to utilise a 'Linear Feedback Shift Register' implementation, it does not exploit large parallel processing. Consequently, the design of a single stage within a systolic array appears to underutilise the optical technology with its potential for large parallel processing. However, an extension of the systolic array architecture is introduced, which effectively cascades the stages of the array by illuminating different areas of the same etalons. Consequently, without increasing the hardware, very long data sequences can be accommodated.

With further development of the technique it may be possible to introduce reconfigurability to achieve improvements in fault tolerance.

REFERENCES

ACKL85 Ackley D.H., Hinton G.E., Sejnowski T.J., "A Learning Algorithm for Boltzmann Machines," Cognitive Science 9, 1985, pp. 147-169.

ARVI82 Arvind, Gostelow K.P., "The U-Interpreter," IEEE Computer, Vol 15, No.2, Feb. 1982, pp. 42-49.

ATHA85 Athale R.A., Szu H.H., Friedlander C.B., "Attentive Associative Memory and its Optical Implementation," in "Optical Processing" (ed. S. Gustafson), Report to SDIO/IST, Dec.1985, pp. 1.1-1.23.

BACK73 Backus J., "Programming languages and closed applicative languages," Proc. ACM Symp. Principles of Programming Languages, ACM, New York, 1973. pp. 71-86.

BACK78 Backus J., "Can Programming be Liberated from the von Neumann Style? A Functional style and its algebra of programs," Comm. ACM, Vol 21, No 8, Aug 1978, pp. 613-641.

BERK75 Berkling K., "Reduction languages for reduction machines," Proc 2nd. Int. Symp. Computer Architecture (Jan. 1975), IEEE, New York, 1975, pp. 133-140.

BLAH83 Blahut R.E.,
"Theory and Practice of Error Control Codes,"
Addison-Wesley, 1983.

CHAN80 Chang B.J.,
"Dichromated Gelatin Holograms and their applications,"
Optical Engineering, Vol. 19, No. 5, 1980.

CLOS75 Close D.H.,
"Holographic Optical Elements,"
Optical Engineering, Vol. 14, No. 5, 1975.

DAGE84 Dagenais M., Sharfin W.P.,
"Cavityless Optical Bistability due to Light Induced Absorption in Cadmium Sulphide,"
Appl. Phys. Lett., 45(3), 1984.

DENN74 Dennis J.B., Misunas D.P.,
"A computer architecture for highly parallel signal processing,"
Proc. 1974 Nat. Computer Conf., AFIPS Press, Arlington, Va, 1974, pp. 402-409.

DENN80 Dennis J.B.,
"Data Flow Supercomputers,"
IEEE Computer, Vol 13, No. 11, Nov. 1980, pp. 48-56.

FAHL79 Fahlman S.E., "NETL: A system for representing and using real-world knowledge," MIT Press, Cambridge, Mass., 1979.

FARH85 Farhat N.H., Psaltis D., Prata A., Paek E., "Optical Implementation of the Hopfield Model," Applied Optics, Vol. 24, No. 10, May 1985 pp. 1469-1475.

FELD85 Feldman J.K.,
"Connections,"
Byte, Vol 10, No. 4, April 1985, pp 277-284.

FISH85 Fisher A.D., Giles C.L.,
"Optical Adaptive Associative Computer Architectures,"
Dig. of Papers IEEE COMPCON, Spring 1985, pp. 342-344.

FISH86 Fisher A.D.,
"A Review of Spatial Light Modulators,"
Tech. Dig. OSA meeting on Optical Computing, Incline Village, Nevada, 1985.

GOLD81 Goldtone A.J., Garmire E.M.,
"On the dynamic response of non-linear Fabry Perot interferometers,"
IEEE Journ. Quant. Electronics, Vol. QE-17, No. 3, 1981.

HILL85 Hillis W.D.,
"The Connection Machine,"
MIT Press, 1985.

HOAR78 Hoare C.A.R.,
"Communicating Sequential Processes,"
Comm. ACM, Vol.21, No. 8, Aug 1978, pp. 666-677.

HOCK84 Hockney R.W.,
"MIMD Computing in the USA - 1984,"
Internal Report, Dept. of Computer Science, University of Reading, U.K.

HOPF82 Hopfield J.J.,
"Neural networks and physical systems with emergent collective computational abilities,"
Proc. Nat. Ac.Sci.(USA), Vol. 79, April 1982, pp. 2554-2558.

INMO84 Inmos Ltd.,
"OCCAM Programming Manual,"
Prentice-Hall, Jan 1984.

JANO85 Janossy I., Tagizadeh M.R.,
Mathew J.G.H., Smith S.D.,
"Thermally Induced Optical Bistability in Thin Film Devices,"
IEEE Journ. Quant. Electronics, Vol. QE-21,
No. 9, Sept. 1985.

JENK83 Jenkins B.K., Strand T.C.,
"Computer-generated holograms for space-variant interconnections in optical logic systems,"
Proc. SPIE Int. Soc. Opt. Eng. (USA),
Vol. 437, 1983.

KUNG79 Kung H.T.,
"Lets Design Algorithms for VLSI Systems,"
Proc. Conf. VLSI Architecture, Design, Fabrication.,
Caltech, Jan 1979 pp. 65-90.

LEE70 Lee W.H.,
"Sampled Fourier Transform Hologram generated by Computer,"
Applied Optics, Vol. 9, No. 3, 1970.

LIU84 Liu K.Y.,
"Architecture for VLSI Design of Reed-Solomon Decoders,"
IEEE Trans. Comp., Vol C-33, No. 2, Feb. 1984, pp. 178-189.

MASS69 Massey J.L.,
"Shift-Register Synthesis and BCH Decoding,"
IEEE Trans. Inf. Th., Vol. IT-15, No. 1, Jan. 1969,
pp. 122-127.

MILL81 Miller A., Miller D.A.B., Smith S.D., "Dynamic Non Linear Optical Processes in Semiconductors," Advances in Physics, 30(6), 1981.

MINS68 Minsky M., Papert S., "Perceptrons," MIT Press, Cambridge Mass., 1968.

NORR84 Norrie C., "Supercomputers for Superproblems: An Architectural Introduction," IEEE Computer, March 1984, pp. 62-74.

PARN85 Parnas D.L., "Software Aspects of Strategic Defense Systems," Communications ACM, Vol. 28, No. 12, Dec. 1985, pp. 1326-1335.

PEYG83 Peyghambarian N., Gibbs H.M., et al, "Observation of Biexcitonic Optical Bistability and Optical Limiting in Cu Cl," Phys. Rev. Lett., 51(18), 1983.

RAND78 Randell B., Lee P.A., Treleaven P.C., "Reliability issues in computing system design," ACM Comp. Surveys, Vol 10, No 2, June 1978.

ROSE62 Rosenblatt F., "Principles of Perceptrons," Spartan, Washington D.C., 1962.

SILV83 da Silva J.G.D., Watson I., "Pseudo-associative store with hardware hashing," IEE Proc., Vol.130, Pt.E, No. 1, Jan 1983, pp. 19-24.

SMIT85 Smith S.D., Janossy I., Mathew J.G.H., Reid J.J.E., Taghizadeh M.R., Tooley F.A.P., Walker A.C., "Non-linear Optical Circuit Elements as Logic Gates for Optical Computers : The First Digital Optical Circuits," Optical Engineering, Vol 24, No. 4, 1985.

TAI82 Tai K., Maloney J.V., Gibbs H.M., "Optical Crosstalk between nearby Optically Bistable Devices on the same etalon," Optical Lett., 7(9), Sept. 1982.

TARN84 Tarng S.S., Gibbs H.M., et al., "Use of a diode laser to observe room temperature, low-power optical bistability in a GaAs-AlGaAs etalon," Appl. Phys. Letters, Vol. 44, No. 4, 1984.

TOOL86 Tooley F.A.P., "Practical Considerations in the construction of Digital Optical Circuits," (1986 - to be published).

TREL82 Treleaven P.C., Brownbridge D.R., and Hopkins R.P., "Data-Driven and Demand-Driven Computer Architecture," ACM Comp. Surv., Vol 14., No. 1, March 1982, pp. 93-143.

TREL83 Treleaven P.C., Mundy D., Lima I.G., "Decentralised Control Flow Programming," Proc. IFIP World Computer Congress, Sept. 1983, pp. 487-492.

TREL84 Treleaven P.C., Lima I.G., "Future Computers: Logic, Data Flow, ...Control Flow?", IEEE Computer, March 1984, pp. 47-57.

TREL86 Treleaven P.C., Refenes A.N., Lees K.J., McCabe S.C., "Computer Architectures for Artificial Intelligence," Internal Technical Report, Dept. of Computer Science, University College London, U.K.

VEGD84 Vegdahl S.R., "A Survey of Proposed Architectures for Execution of Functional Languages," IEEE Trans. Computers, Vol. C-33, No. 12, Dec 1984, pp. 1050-1071.

WALK86 Walker A.C., "The Application of Bistable Optical Logic Gate Arrays to All-Optical Digital Parallel Processing," (1986 - to be published).

WANG85 Wang C.P., Smith P.L., "Charged-large-array-flexible mirror," Applied Optics, Vol 24, No 12, June 1985, pp. 1838-1843.

WATS82 Watson I, Gurd J.R., "A Practical Data Flow Computer," IEEE Computer, Vol.15, No.2, Feb 1982, pp. 51-57.

WHER84 Wherrett B.S., "Fabry Perot Bistable Cavity Optimisation on Reflection," IEEE Journ. Quant. Electronics, Vol. QE-10, No. 6, 1984.

WHER85 Wherrett B.S., "All Optical Computation - A Parallel Integrator based upon a Single Full Adder," Optics Communications, Vol. 56, No. 2, 1985.

WHER86 Wherrett B.S., Hutchings D., Russell D., "Optically Bistable Interference Filters : Optimisation Considerations," (1986) to be published.

AQM003091AA01

SECTION B: RECOMMENDATIONS FOR FUTURE PROGRAMS

1. INTRODUCTION

Section B of this report identifies suitable topics for future research programs to be performed by Ferranti Computer Systems Limited. These topics are discussed to indicate the potential benefits to be obtained.

Each topic does not imply an individual work program. It is likely that related topics will be combined to form a number of work programs, according to needs and objectives.

1.1 RESEARCH PHILOSOPHY AND ORGANISATION

An effective program relies on extensive interaction to equate the device characteristics to the various system requirements. The philosophy adopted during the initial phase of the research program (Section A, 1.2) was successfully implemented to obtain clear objectives and conclusions. An applications driven philosophy is therefore proposed.

2. PROPOSED TOPICS

2.1 APPLICATIONS AND ARCHITECTURES

No single architecture is optimally suited to meet all applications. Equally it would appear the architectures that we propose for optical general purpose computing are still under development. However, the most significant general conclusion is that an optical parallel computer should comprise many ultra-simple processing elements, with a mechanism for dynamic interconnection between them. Of the five architectural styles classified and considered during our investigation, three types have been identified for further consideration. These are discussed below:

2.1.1 Systolic Arrays

These are generally utilised for fixed, regular functions which are realized as hardware implementations of specific algorithms. The architecture style has been historically developed to exploit the limitations of VLSI planar structures for very high performance processing. It is likely that the concept of systolic processing may be extended to exploit three-dimensional structures.

This architecture is favoured for high performance "computation intensive" functions, but is generally limited to applications which can be expressed in terms of highly regular algorithms.

For suitable applications it is likely that optical and electro-optical technologies can further enhance performance. However, the following areas need to be addressed if criteria for use and optimisation are to be understood:

- Assessment of suitable topologies and element configurations which could maximise potential within the constraints of a given optical technology.

- Investigation into the implications of optical techniques on fault tolerance capabilities in systolic array implementations.

- Identification and development of basic optical "building block" devices required to implement elements of the array.

- Investigation into suitable number representations and arithmetic techniques for systolic array applications.

2.1.2 Data Flow Architectures

A key feature of this architecture is that it supports programming using a Functional Language. Parallelism is implicit (i.e. parallel processes do not need to be specified explicitly within the written program), and the result of executing a program is mathematically equivalent to the program definition. This may lead to simplified program generation and, perhaps more importantly, software verification. It would appear that a Data Flow Architecture requires global interconnections (with dynamic adjustment), and very simple node processing circuits in order to successfully achieve an optical implementation. It is an architecture which demands large parallelism in a program (or set of programs) in order to achieve the maximum use of physical resources. Similarly the desirable system properties outlined in Section A can be provided by careful design. Consequently Data Flow architectures are recommended for further consideration, and the following areas need to be addressed:

- Investigation and development of simple node processors utilising optical or electro-optical technologies.

- Investigation and development of Data Flow architectures: initially based on the static model but ulcimately on the dynamic tagged model, considering the detailed implications of an optical implementation.

- Investigation and development of suitable associative memory mechanisms for matching data tokens, as required to implement a dynamic tagged data flow model.

- Investigation and development of techniques for achieving global and dynamic interconnections.

- Investigation and identification of number representations and arithmetic techniques suitable for general purpose applications in a Data Flow machine.

- Identification and development of basic optical "building block" devices required to implement the primitive functions, provide selection and control, and data interfaces.

2.1.3 Massive Parallel Architectures

Neural Networks and the Connection Machine are examples of these types of architectures, which support nearest-neighbour search and semantic network operations, for example. Pattern recognition, associative memory and error correction functions illustrate the potential application areas for architectures, and typify the classes of problem for which the von Neumann Computer seems ill-suited.

Successful implementation of a massively parallel "connectionist" architecture using optical and electro-optical devices appears to rely upon the ability to make global and dynamic interconnections between processing nodes. Similarly the processing elements need to be designed as an extremely simple function for economic utilisation of the optical technologies. Consequently the following areas need to be addressed:

- Investigation and development of an optimum configuration for a processing element utilising optical devices.

- Investigation and development of a suitable system hardware organisation.

- Investigation and development of techniques for global and dynamic interconnections.

- Identification and development of basic optical "building block" devices required to implement the primitive functions, provide selection and control, and data interfaces.

- Investigation and identification of number representations and arithmetic techniques, as appropriate.

2.2 TECHNOLOGIES

During the initial phase a comprehensive assessment was carried out to determine performance requirements and potential for non-linear interferometers and supporting devices and technologies. An important achievement for this program was to establish engineering specifications which detailed those parameters and characteristics which would influence design.

By scrutinising these specifications the essential areas for optimisation could easily be identified, together with a clear perspective of tradeoff mechanisms. As each of the target specifications was derived using current knowledge of potential and practically achievable performance requirements, after considering the system functions and relating to all components and technologies which make up the circuits, then the shortcomings for any device can be easily quantified in terms of system performance. Consequently, meaningful comparisons can be achieved.

The following topics have been identified for future programs, and these are discussed below:

2.2.1 Etalon-Based Technology

Continued research and development for improvements in performance, for non-linear interferometer devices, supporting device technologies and their interconnection, is essential if rapid evolvement of optical computing systems is intended. As each improvement is realised, and the inevitable tradeoffs are quantified, then new targets must be established for overall optimisation. The procedure is iterative and highly interactive, and includes the following activities.

- Review of current capabilities; technology characterisation and specification.

- Review of application techniques and performance criteria.

- Identification and development of basic optical "building block" devices.

- Assessment of target specifications.

- Design of demonstrator functions.

- Practical evaluation.

2.2.2 Planar Based Technology

It is envisaged that mixed etalon-planar configurations will optimise circuit and system performance and implementation. It is anticipated that specific requirements for achievement and improvements will be identified by carrying out a similar investigation and assessment for the planar technology, as has been performed for the etalon technology. This would include the following activities:

- Identification and review of candidate technologies for planar construction.

- Assessment and characterisation to establish fundamental and practical limitations.

- Investigation and development of applications and architectures which exploit optical planar technologies.

- Investigation and development of basic optical elements, interconnection techniques, steering, control and interface devices required for implementation.

- Assessment and development of engineering specifications for current performance and projected requirements.

- Development of demonstrator functions.

- Evaluation of performance factors to determine tradeoffs for optimisation.

2.2.3 Power Control

For optical systems employing power modulation for describing data and control, generally the penalty suffered is low efficiency in the generation and use of power. An inherent problem for optical systems of this kind is that natural techniques for optical power regulation do not exist. By analogy, electronic systems can easily monitor voltage and current, and the same voltage or current can be arranged as negative feedback (subtraction) to hold these parameters within limits.

For the optical systems considered within the initial research program, and for traditional and innovative liner transformation processes using the Fourier Transformation properties of lenses and holograms, the control and regulation of power sources becomes critical.

An investigation and development of modulation and feedback mechanisms is recommended.

END

10-86

DTIC